


RESOURCE-RELATED RESEARCH
COMPUTERS AND CHEMISTRY
(RR-00612 ANNUAL REPORT)

Submitted to
BIOTECHNOLOGY RESOURCES BRANCH
OF THE
NATIONAL INSTITUTES OF HEALTH


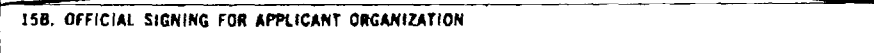
May, 1975

COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY

SECTION I

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE PUBLIC HEALTH SERVICE		REVIEW GROUP SSS	TYPE 5	PROGRAM R24	GRANT NUMBER (Insert) RR00612-05A	
TOTAL PROJECT PERIOD From: 05/01/74 Through: 04/30/77						
REQUESTED BUDGET PERIOD From: 05/01/74 Through: 07/31/75						
1. TITLE RESOURCE RELATED RESEARCH - COMPUTERS AND CHEMISTRY						
2A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR (Name and Address, Street, City, State, Zip Code) DJERASSI, CARL STANFORD UNIVERSITY DEPT OF CHEMISTRY STANFORD, CALIF 94305			4. APPLICANT ORGANIZATION (Name and Address-Street, City, State, Zip) STANFORD UNIVERSITY STANFORD, CALIF 94305			
2B. DEGREE PHD	2C. SOCIAL SECURITY NO. 		5. PHS ACCOUNT NUMBER 1941156365A1			
2D. DEPARTMENT, SERVICE, LABORATORY OR EQUIVALENT CHEMISTRY			6. TITLE AND ADDRESS OF OFFICIAL IN BUSINESS OFFICE OF APPLICANT ORGANIZATION DEPUTY V P FOR BUSINESS & FIN STANFORD UNIVERSITY STANFORD, CALIF 94305			
2E. MAJOR SUBDIVISION SCH OF HUMANITIES AND SCIENCES						
3. ORGANIZATIONAL COMPONENT TO RECEIVE CREDIT FOR INSTITUTIONAL GRANT PURPOSES 20 OTHER						
7. RESEARCH INVOLVING HUMAN SUBJECTS (See Instructions) <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes APPROVED _____ Date _____			8. INVENTIONS (See Instructions) <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes-not previously reported <input type="checkbox"/> Yes-previously reported			
9. PERFORMANCE SITE(S) Stanford University Department of Chemistry Computer Science Department Department of Genetics Stanford, CA 94305			TELEPHONE INFORMATION			
			11A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR (Item 2a)		Area code	Tele. No.
			11B. Name of business official (Item 6)			
			11C. Name and title of administrative official (Item 15b)			
10. DIRECT COSTS REQUESTED FOR BUDGET PERIOD \$240,962			11B. Name of business official (Item 6) K. D. Creighton			
12A. CONGRESSIONAL DISTRICT OF APPLICANT ORGANIZATION SHOWN IN ITEM 4 Twelfth			11C. Name and title of administrative official (Item 15b) D'Ann Downey, Assistant Sponsored Projects Officer			
			12B. COUNTY OF APPLICANT ORGANIZATION SHOWN IN ITEM 4 Santa Clara			
13. DO NOT USE THIS SPACE						

14. CERTIFICATION AND ACCEPTANCE. We, the undersigned, certify that the statements herein are true and complete to the best of our knowledge, and as to any grant awarded, the obligation to comply with Public Health Service terms and conditions in effect at the time of the award.

15. SIGNATURES (Signatures required on original copy only. Use ink. "Per" signatures not acceptable.)	15A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR 	DATE
	15B. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION 	DATE

II. RESOURCE OPERATIONS

A. DESCRIPTION OF PROGRESS

B. SUMMARY OF RESOURCE USAGE

C. RESOURCE RELATED RESEARCH EQUIPMENT LIST

D. SUMMARY OF PUBLICATIONS

A. DESCRIPTION OF PROGRESS

OVERVIEW

In the first twelve months of this fifteen-month grant period, the DENDRAL programs and the GC/MS data system have moved significantly forward under NIH funding, even though it was partly a time of transition from one computer system to another. This report of progress is organized in three parts, corresponding to the three specific aims of our December, 1973, proposal: (PART 1) Enhancing the power of the mass spectrometry resource, (PART 2) Developing performance and theory formation programs, and (PART 3) Applying the computer programs and instrumentation to biomedically relevant structure elucidation problems.

The highlight of the period since May 1, 1974, was the project's move to the interactive computing environment of the NIH-funded SUMEX-AIM facility from the batch computing environment of the Stanford Computation Center. Because of this, many scientists outside this university have been able to use the DENDRAL computer programs for their own research. Also, the programs themselves grew in power and scope, and we opened new vistas for collaboration with other research groups. We have been able to make the programs more conversational and thus more helpful to the chemists and biochemists for whom they were developed. Outside users in other research groups also have in SUMEX an easy mechanism for trying out the DENDRAL programs on their own structure elucidation problems. Finally, we have a mechanism for looking at subroutines developed by other research groups in the context of our own programs -- and have incorporated subroutines written, for example, by T. Wipke and by R. Feldmann, into our procedures. The programs and their development are discussed in Part 2, below.

The DENDRAL project, one of the major users of the SUMEX-AIM computer facility, has been forming its own community of remote users. This "exodendral" community has already provided valuable contributions to program development and both the community and contributions are expected to grow at an increased rate. As an example, for the last month for which figures are available (March 1975), the number of CPU hours used by exodendral persons amounted to at least ten percent of the CPU hours used by the entire DENDRAL project. In the last month alone, one new exodendral account representing at least three users has been added to the system, and another four exodendral users have been invited to begin their usage via various "guest" accounts.

Another milestone in this period was the delivery of the PDP-11/45 computer, and successful transfer of data acquisition and reduction programs into that computer. This has provided a stand-alone environment for our mass spectrometer/computer system

in which development, experimentation and routine use of the system is much more simple, reliable and efficient than previously. We have made excellent progress in fulfilling the goals of combined gas chromatography/high resolution mass spectrometry (see Part 1, below).

Our programs are receiving heavy use from local users and outside users who are investigating mass spectrometry problems for a variety of different compound classes. In addition, new program developments have extended the scope of biomedical structure elucidation problems for which we can provide some computer assistance. Local users include members of Professor Djerassi's group, other chemistry department persons and research groups at the Stanford Medical School. We have recently begun the process of building a community of outside users who can access our programs at SUMEX via TYMNET or ARPANET. Several research groups have expressed considerable interest; we have demonstrated and explained the programs to several groups and we are currently arranging more demonstrations and assisting other people in learning to use SUMEX and the programs from their own laboratories. These applications are discussed in detail in Part 3, below.

1 PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE

1.1 Introduction

Our grant proposal requested funds for significant upgrading of our capabilities in mass spectrometry. The goals of this upgrading were to provide routine high resolution mass spectrometry (HRMS), combined gas chromatography/low resolution mass spectrometry (GC/LRMS) and to develop a combined gas chromatography/high resolution mass spectrometry (GC/HRMS) facility. In addition, this would provide the capability for new experiments in the detection and utilization of data on metastable ions. These capabilities would then be available as required for application to our wider goal, solution of biomedical structure elucidation problems of community of researchers.

The upgrading included several items of hardware and software development, as follows: 1) Acquire stand-alone computer support for the mass spectrometer because existing facilities were inadequate and very expensive; 2) convert existing software, written in the PL/ACME language into FORTRAN so that it would run on the new system; 3) develop new software as required for the demanding task of GC/HRMS; 4) provide hardware and software for semi-automatic acquisition of data on metastable ions. The initial development phase of this upgrading included performance tests to determine the capabilities and limitations of the GC/HRMS system to define the scope of problems to which it can be applied.

The present status of this effort is that the computer system has been purchased, installed and is operating. The software has been converted and is operational. The GC/HRMS system is in the trial stage and is working. Future developments include significant improvements in the software to provide more routine and reliable GC/HRMS operation and to provide better information to the operator on instrument performance and to the chemist on the characteristics of his data. The metastable ion work has been deferred until now because of the more pressing demands of the GC/HRMS system, but work can now begin on this aspect of our research.

We presently are in a position to provide routine LRMS and HRMS support for our chemical research and program development. The GC/HRMS system is working well enough to commence study of real problems. The above developments and future goals are summarized in detail in the subsequent sections.

1.2 Hardware Acquisition and Development

We have, in the mass spectrometry laboratory, two mass spectrometers which were connected to our previous computer system (ACME), the Varian-MAT 711 mass spectrometer, and the AEI MS-9, both high resolution mass spectrometers. We have concentrated our efforts to this point on development of the 711/computer system because this (much more modern) instrument is the spectrometer of choice for the GC/HRMS experiments. It was already equipped with the gas chromatograph and GC/LRMS work was already routine as far as the mass spectrometer system was concerned. We were granted some money for minor upgrading of the MS-9 so that it could relieve some of the burden of routine HRMS analysis from the 711 (see Future Developments).

At the time the grant was awarded, we were essentially without computer support in the mass spectrometry laboratory. Interim funds were provided to permit us to connect to the ACME system running on a different computer system. This connection was made and permitted us to obtain some HRMS data while purchase and installation of the stand-alone system was completed. The interim ACME system was even less effective than in its previous environment, and did not permit GC/MS experiments due to slow response time.

We concurred with the study section's recommendation that stand-alone computer support be provided for efficiency and long-term cost effectiveness, and that such support be a PDP 11/45 or equivalent as the machine with the capabilities to handle the heavy data burden imposed by GC/HRMS. We were able to adjust our first year budget to allow purchase of this computer. It was ordered on Feb. 11, 1974 as a contingent order (contingent on award of the grant). The firm order was placed March 18, 1974. It was actually delivered on August 2, 1974. Together with the Digital Equipment Corp. (DEC) PDP-11/45, we obtained a disk system from Systems Industries because it was considerably less expensive than the comparable DEC device.

A diagram of the current hardware system is shown in Figure 1. Note that this system is interfaced to the mass spectrometer through a previously existing PDP-11/20. There are two important reasons for this configuration. The 11/20 contained the necessary hardware extensions for computer control of the mass spectrometer under the old ACME system. This drastically reduced the mass spectrometer/computer interface problems. The 11/20 acts as a buffer between the mass spectrometer and the 11/45, thus freeing the 11/45 for computations during the course of data acquisition. This is an important element of future foreground/background processing.

1.3 Software Development

Conversion of existing PL/ACME programs to FORTRAN was begun on the award of the grant. The delays in delivery of the 11/45 system caused delays in this development because no machine was available for certain tests of the programs, and of course, no work on new mass spectral data could be done until the mass spectrometer and computer system became operational. Conversion of these algorithms also included many system software developments to ensure that previously batch processing programs could function in a real-time environment under the requirements of GC/HRMS operation. This development included not only improvements and extensions to existing algorithms, but building a file management system for facile logging and storage of spectra with the ability for simple recall to examine or recompute old data, and a diverse package of debugging, display and plotting and mass spectrometer evaluation programs.

Because we view GC/HRMS as the most important new capability of our mass spectrometer/computer work, the requirements of GC/HRMS have guided development of the software system. These requirements include continuous automatic monitoring of instrument performance to avoid wasting time collecting poor or erroneous data. Because we have chosen to approach GC/HRMS with an electrical recording system, as opposed to photographic, we are able to monitor the instrument continuously, both during initial setup and during the course of the GC/HRMS experiment. Major sections of the software and how they interact among one another are summarized below.

1.4 Software Architecture

Figures 2 and 3 show the various software configurations possible within the GC/HRMS system. The data paths and options available in obtaining reference spectra for instrument diagnostics and calibration are illustrated in Figure 2. Successful operation of the mass spectrometer depends on successful setup and verification of the performance of the spectrometer. The various routines outlined in the Figure permit the operator to acquire data and examine it during each stage of

the subsequent reduction. A typical run might be (using the REFRUN command processor for control of the system) to acquire a spectrum (RRØMOD), reduce the scan data to peak times and areas (RRØRED) while saving the raw data on magnetic tape for future reference. Time to mass conversion and display to the operator (on a CRT display) are automatic and provide both the results and diagnostics (RRØREP). These data may be examined further, e.g., via spectrum bar plots, peak profile examination.

The data paths and options available to the operator for collection of high resolution mass spectral data (whether for single samples or for GC/HRMS) are summarized in Figure 3. The various processors summarized in the Figure have a simply-stated but computationally difficult task; to acquire and quickly reduce a spectrum to masses, elemental compositions and intensities. The task is completely automatic. It is based on information about the particular setup of the mass spectrometer (scan time, duration, reference ions, etc.), determined during the reference ions, etc.), determined during the above procedures. Again, raw data are saved for future perusal and the operator can examine the data at any stage during data reduction. Diagnostic and instrument performance information are available after each scan to monitor continuously the status of the mass spectrometer. In normal use, the process runs without interference. Completely reduced data, for the normal spectrum, are available within 2-3 seconds after completion of the scan. This is so much better turn-around than the previous ACME system that it has opened new possibilities for data acquisition and reduction. For example, in cyclic mode, where spectra are acquired repetitively, the computer system can easily keep up with the mass spectrometer. Diagnostics can point out instrument problems before sample is wasted in another run.

1.5 System Philosophy

The underlying philosophy governing the design and development of the software system is dominated by three considerations. First, the operator-data system interface must be flexible enough to meet the changing and often times novel demands required by the experimental nature of the GC/HRMS procedures. While predefined operational sequences are essential for production processing, such sequences must not be so rigidly defined that deviations cannot be made to accommodate experimental modes of instrument operation. Second, the system must maintain its integrity under severe environmental conditions. Unforeseen and often uncontrollable conditions can cause catastrophic hardware failure. The filing system must be made immune to contamination by such occurrences. Third, the overall system structure must be amenable to organized software growth and expansion. It must be easy to add facilities and to implement and evaluate experimental algorithms and heuristics.

1.6 System Flexibility

Due to the dual role of the system as both a production instrument (at this point HRMS) and an experimental instrument, (GC/HRMS and metastable work), the operator-data system interface must provide both a convenient means of executing often utilized operational sequences as well as a flexible means of exploring sequences amenable to the experimental work. Towards this end each major system program consists of a resident command driven interpreter. This interpreter accepts a two letter keyword command from the operator and then invokes an overlaid semantic routine. If an unknown command is entered by the operator, facilities exist for refreshing the operator's memory of which commands are appropriate under the circumstances. Semantic routines may interrogate the operator directly or may utilize default information contained within a disk file. Such a simple structure provides for easy expansion of system facilities while also providing for explicit control of the sequence of operations by the operator.

In addition to the control structure within the PDP 11/45, facilities also exist for controlling the PDP 11/20 directly. Each process which can be loaded into the PDP 11/20 has a finite number of distinct states. Operator commands exist to cause transitions between each of these states. Thus, it is impossible for the PDP 11/20 processor to get 'stuck' waiting for an event which will never occur.

1.7 System Integrity

The minute quantities of certain samples which have been submitted for analysis prohibit the re-running of any experiments associated with these samples. The system operates in a somewhat hostile environment. The physical laboratory environment dictates that the computer system be located in close proximity to the GC/MS instrument. The instrument can cause severe electromagnetic disturbances (sparks within the source, high voltage shut down, etc.) which can bring down either the entire data system or portions of the system. Static electric discharges from the operator through the system console have also resulted in catastrophic consequences for the data system. These occurrences are quite unpredictable from the software point of view and are difficult to alleviate in the physical environment. Therefore, the software must file data as soon as it is acquired in order that in the event of system failure any data gathered up to that point is maintained intact. It is for this reason that the thresholded data is logged directly onto magnetic tape by the PDP 11/20 processor. It is for this reason also that the reduced data filing mechanism insures that the last data block of a scan is actually written out onto the disk and not left in a DOS system buffer.

Some of these topics are amplified in the subsequent sections where several features of the system are described in more detail.

1.8 Operating System

[Between the time this section was drafted and finished, the conversion to the DOS 9 operating system was made.]

DOS version 8 is the operating system currently in use. However, we have converted the software to DOS 9 and are awaiting the installation of the IMS disk system. The major mandate for this conversion is the vastly improved overlay system offered by the new operating system. Overlaid files are maintained as a single, contiguous file on disk as opposed to the DOS 8 method of maintaining a separate linked file for each overlay. The DOS 8 strategy demands that a linked file be opened, read, and closed for each overlay load. DOS 9 allows an overlay to be loaded with a single disk read. Also the DOS 9 overlay facility provides for tree structuring process which was completely absent from DOS 8. Considering that the current version of the system has 17 overlays the importance of efficient overlay loading is obvious. In addition to these factors, DOS 9 provides us with batch processing facilities which make it much easier to do system generation, archive data, etc. The decision to use DOS 9 was a major factor in switching from the System Industries, non-DEC compatible drives to the IMS fully software compatible drives.

1.9 GC/HRMS Software System

On top of DOS the GC/HRMS system has been constructed. This system has both real time as well as non-real time facilities. The real time facilities include: 1) the acquisition and reduction of a reference run to calibrate the instrument and 2) the acquisition and reduction of a sample run. Both processes must install the acquisition routines into the PDP 11/20.

The non-real time facilities include the ability to re-examine reduced data files, to re-run the reduction processes from the back-up medium, to communicate to the PDP 10 or other processors the results of composition matching for use in other processing, for example, mass spectra for MOLION, PLANNER or INTSUM (see Part 2, below).

1.10 General Data Acquisition Procedure

The actual data acquisition procedure is accomplished in two steps. The first step involves calibrating the instrument for the particular set up of the MAT 711 and the second step involves the actual analysis of the sample.

The program REFRUN provides the calibration facilities for the instrument. The operator runs the program and informs the system of the sampling rate, the direction of the mass scan, routing of displays. The final tweaking of the MAT 711 is performed and a scan command is performed. The PDP 11/45

commands the previously loaded process in the PDP 11/20 to start the scan. The 11/20 checks that the MAT 711 interface is set up properly and then initials the scan. When the scan completes the PDP 11/20 is signaled through the mass spec control interface and it compiles a spectrum trailer which it tacks onto the end of the peak profile data it is logging and sending to the 11/45. As the 11/20 feeds these data into the 45, it is crunched down into time intensity pairs. When the spectrum trailer comes through the 11/45 invokes the mass computation algorithm which attempts to locate the prominent landmarks in the PFK spectrum. If this process is successful, a display is produced showing the model peak profile, the resolution as a function of mass, and the projection errors for calibration of the mass/time curve of the instrument. In addition, signal and noise information is displayed. If this process is repeatable (in the sense of taking 2 or 3 scans which yield essentially the same results) and performance at a sufficiently high level, then the reduced data is filed for later use by the sample analysis routines.

The program SAMRUN provides the sample analysis facilities. It is run after a successful REFRUN. The information about the time to mass conversion from the preceding refrun is used to perform the time to mass conversion for the sample spectra. The rapid, automatic nature of this procedure was mentioned above.

1.11 Filing Systems

The filing system can be roughly divided into three components. First, when a spectrum is acquired from the MAT 711 it is thresholded and background removal is performed to produce what is called peak profile data. This data is logged immediately onto magnetic tape by the PDP 11/20. Second, as data is being reduced into mass amplitude pairs by REFRUN or SAMRUN it is filed onto disk for easy retrieval for later examination. The format of the reduced files for SAMRUN and REFRUN is slightly different due to the fact that a refrun contains only one spectrum while a sample may have any number of spectra. Third, the system maintains files which contain control information, spooled hardcopy information and composition output to be transferred to the PDP 10 for further analysis.

1.12 Buffering

Buffering is a central issue in the system. Due to the uneven distribution of data, high data rates, slack periods, it is desirable to provide a large amount of buffering between the instrument itself and the reduction processes. It is the case that data from one spectrum can be reduced while another spectrum is being acquired. Currently the PDP 11/20 has sufficient buffer capacity to hold almost a complete spectrum of a sample. The 11/45 will soon have the capability to run the peak profile to intensity-time reduction process concurrently with the time

intensity to mass amplitude conversion process or the display processes. Such concurrence is another side effect of the operation under DOS 9 due to the ability to tree structure overlays.

This software system permits a great deal of flexibility in operation of the system. Where time between scans permits (a few seconds) the HRMS data can be reduced completely to accurate masses and intensities, and feedback provided to the operator on the quality of the scan. This output is the data used to determine the quality of the spectral data. One can disregard scans which are poor and know when one is of high quality. Alternatively (and in addition to the above mode), data (peak profiles and intensities) can be spooled onto tape for later processing. The operator can choose to print out results immediately for critical samples, or defer final output until later while additional data are being collected. An archival system provides the facility for storing and retrieving old spectral data for review or reanalysis.

1.13 Present Status and Performance Tests

As mentioned previously, the system is operational now and is in routine use for HRMS and in experimental use for GC/HRMS. New developments (see below) and routine use are proceeding in concert. We maintain the previous version of the programs for routine use while work continues on a separate version to add improvements, remove program "bugs" and so forth. We are rapidly proceeding toward a system which is relatively "crash proof" in spite of what the mass spectrometer might do and in spite of abnormalities which might appear in the data. Such a system is critical to ensure the integrity of data collected during a long GC/HRMS run.

We have been running many performance tests on the GC/HRMS system to determine what problems arise when the mass spectrometer and computer system are pushed to their utmost, and to determine the sensitivity in terms of sample size of the GC/MS combination. In several instances, additional programming had to be done to cope with the demands of GC/HRMS. These additions are largely complete now (e.g., allocation of extra memory and disk space for REFRUN when the GC is at high temperature and considerable numbers of ions from GC column bleed are present in addition to the internal mass standard, perfluorokerosene) and we have turned our attention to measuring performance samples.

In performing sensitivity tests we can always make the system look good by choosing samples which have characteristics such that excellent mass spectra can be obtained on minute amounts of material. We have not done this. We have chosen samples which are representative of the types of material that are the focus of our current chemical research interest. For the simpler case of fatty acid methyl esters, we can obtain, in 8-10 sec. per decade in mass scans, HRMS displaying ions over a

dynamic range of 100:1 with about 1 microgram of material per component. For the harder case of free sterols, such as cholesterol, 2 micrograms per component yields the same performance (such sterols have many more ions over a wider mass range, thus requiring more material for the same dynamic range of ion intensities). We do not claim that this sensitivity is the ultimate achievable. It is certainly sufficient for many of our problems. It will be insufficient for those problems where there are components over a wide range of concentration. Because the GC column cannot be overloaded too severely for the major components it is difficult to increase greatly the concentration of the minor components. Additional chemical or physical separations can solve some problems of this type. But with this sensitivity we can get much useful work done as we progress with improvements to our techniques (see Future Developments).

Current applications of the mass spectrometer/computer system to biomedical structure elucidation problems are summarized in Part 3 of this report.

1.14 Future Developments

The second year of our grant has several specific goals for the mass spectrometer/computer system. These goals, outlined below, will improve the performance and reliability of the current system to ensure the integrity of results on precious samples. This year will also be taken up with implementing the other hardware and software items mentioned in our original proposal (metastable ion work, MS-9 hook up, multiplet detection and resolution) now that the primary goal of GC/HRMS is well in hand.

1.14.1 Hardware

The following are the important goals in hardware development, necessary to fulfill our research objectives.

A) Installation and testing of the hardware for control of the mass spectrometer for semi-automatic acquisition of data on metastable ions. This hardware, including circuitry to interface the metastable scanning system of the Varian MAT 711 to the 11/45 system, and a high precision A/D converter, will permit semi-automatic detection and analysis of metastable ions which relate a "daughter" fragment ion to its progenitors, "parents". It will allow us to explore the inverse relationship, determination of all daughter ions from a given parent ion by simultaneous variation of two of the three fields in the instrument, accelerating voltage, electrostatic analyzer voltage and magnetic field. The feasibility of this technique has apparently been demonstrated by Lacey and McDonald in Australia.

B) Reconnection of the MS-9 to the new 11/45 data system.

There is a relatively minor amount of work which must be done to enable us to acquire data from the MS-9 under the new instrument control structure of the 11/45 computer system. This will enable us to divert routine samples to the MS-9 for analysis and permit us to devote more time on the 711 to the more difficult task of GC/HRMS.

C) Connection of a plotter to the computer system. An existing Cal-Comp plotter will be connected to the system so that hard copy of graphical (e.g., mass spectra, instrument performance curves) output can be obtained. Presently only CRT output of this information is available.

1.14.2 Software

With the hardware largely installed and functioning, the software (the actual programs which acquire, manipulate, reduce and output the mass spectral data), requires the greatest attention in the coming year. There are several steps to be taken which will improve the capabilities of the GC/MS system in general, GC/HRMS in particular. These are outlined below.

A) Improve data reduction facilities for scanning at lower resolving powers. HRMS data are usually collected at a resolving power sufficient to separate many (but never all) of the possible multiplets of ions possessing the same nominal atomic mass but differing in elemental composition. However, for maximum sensitivity, with relatively little degradation in data quality, one would like to run the mass spectrometer at lower resolving powers. This increases the the likelihood of overlapping peaks which are viewed as single peaks according to our present data acquisition system. Our proposal discussed ways to use both mathematical routines and chemical intelligence to help solve this problem, thus providing effectively higher resolution via data processing, at high sensitivity. We are just implementing the first phase of this approach, to resolve quickly those overlapping peaks whose profiles are well-defined. We view these developments as essential to the success of GC/HRMS because of the improved sensitivity.

B) Better inter-computer communication. We are currently implementing better inter-machine communication between the 11/20 and 11/45 computers (see Fig. 1). This will improve the reliability of the system by allowing "clean" (i.e., restartable) recovery from error conditions in the mass spectrometer, in the input data stream or during data reduction between scans.

C) Implement a GC/LRMS system. Because of our focus on GC/HRMS, the ability to handle efficiently GC/LRMS data has been neglected. We will remedy this situation in year two because some of our problems do not require HRMS data and rapid presentation of LRMS data to the chemist in graphical form will be very useful. The LRMS system will be relatively simple to implement because almost all of the same data acquisition and reduction programs written for HRMS data can be utilized.

D) Software for metastable ion analysis. Routines must be written to enable facile calibration of the mass spectrometer operating in metastable ion mode, and to allow subsequent reduction of data. Existing control and data acquisition software will be used for initiating the metastable ion mode.

1.14.3 Summary

As the above hardware and software improvements are being made we will continue evaluation of the GC/HRMS system in parallel with its actual application to real problems. GC/HRMS is a relatively new and difficult technique for routine application. In order to use it effectively, we will have to exert some effort toward determining and optimizing the performance of the many elements of the system, the GC, the MS, and the computer hardware and software.

2 PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS

TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

2.1 Introduction

The Heuristic DENDRAL computer programs assist with structure elucidation problems by helping interpret mass spectra and helping generate structures that are consistent with the interpretations. The Meta-DENDRAL programs assist with rule formation problems in cases where the rules of mass spectrometry are not known.

In this section we describe our progress on the computer programs. Generally speaking, it has been a productive year because of the interactive computing environment provided by the SUMEX facility. Not only are we able to develop programs much more rapidly than in a batch environment, but we are able to make the programs themselves highly interactive, and thus more useful.

All programs have been transferred to the SUMEX machine and most have been considerably improved from their previous versions. The CONGEN and PLANNER programs, in particular, have been improved substantially because these two were thought to offer the most to scientists with structure elucidation problems. Two new programs were developed in this period: CLEANUP and MOLION. The CLEANUP program helps separate the mass spectra of individual components from a GC/MS analysis, and eliminates the background due to GC column "bleed". The MOLION program determines the mass and empirical formula of the whole molecule from its mass spectrum, without prior knowledge of any of the features of the molecule. Both of these new programs solve major

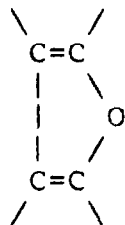
problems that we had previously assumed were already solved before a scientist used the DENDRAL programs.

2.2 CONGEN.

The CONGEN[48,53] program represents a significant extension of a program which has developed over the last several years, the cyclic structure generator[40,41]. The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by 1) allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectroscopy, chemical degradation, method of isolation, etc.) and 2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the program allows interaction at every stage; based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of final structures.

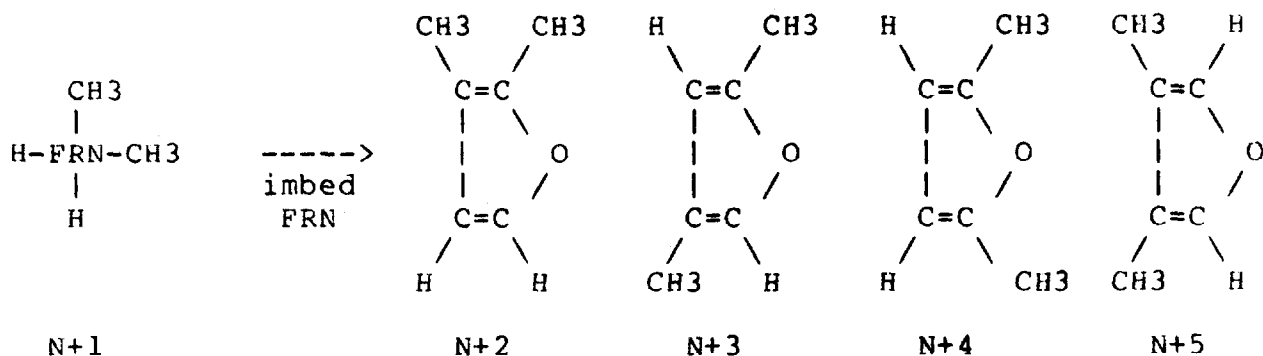
CONGEN fits with the other DENDRAL programs as a "backstop" solution to structure elucidation problems. If the mass spectrum of an unknown compound is available, then CLEANUP and MOLION could be used, but if the general class of the compound is not known, PLANNER has no starting point from which to work. In such cases, structural information can be extracted manually from the spectrum and given to CONGEN for analysis. Because CONGEN makes no assumptions about the source of this information, other spectroscopic or chemical techniques may be used to supply supplemental data.

At the heart of CONGEN are two algorithms whose accuracy has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm[31,37,40,41] is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. For example, in a compound known to contain a furan ring, the quadrivalent superatom FRN might be defined, which has the structure N. Here, the bonds with



N

unspecified termini represent available bonds to hydrogen or other atoms or superatoms. Because the structure generation algorithm can produce only structures in which the superatoms appear as single atoms (we refer to these as intermediate structures), a second procedure, the imbedding algorithm[48,53] is needed to expand the superatoms to their full chemical identities. For example, N+1 is a simple intermediate structure



which might be produced by the structure generator. The imbedding of FRN yields four final structures, N+2-N+5. The output of the imbedder is exhaustive and, in a limited technical sense, free from duplication. But when a list of intermediate structures undergoes imbedding, duplicates can arise. Thus the imbedder is also equipped to post-test such lists for duplicates and retain only unique structures.

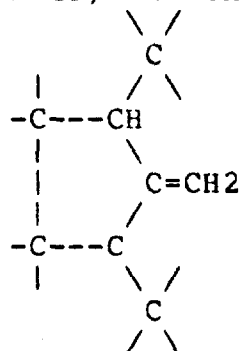
These two routines give the chemist the ability to construct structures from a given set of molecular "building blocks" which may be atoms or larger fragments. By itself, this capacity is of limited utility because the number of final structures can be overwhelming in many cases. Usually, the chemist has additional information (if only some general rules about chemical stability, which the program has no concept of) that can be used to limit the number of structural possibilities. For example, he may know that because of a compound's stability, it cannot contain a peroxide linkage (O-O) and thus the programs need not consider such structures when there are two or more oxygens in the "building block" list. During the past year, a substantial amount of effort has been devoted to modifying these two basic procedures, particularly the structure generation algorithm, to accept a variety of other structural information (constraints), using it as efficiently as possible to prune the list of structural possibilities.

Specifically, there are six types of constraints that we have implemented. GOODLIST and BADLIST are used respectively to require and forbid the presence of user-specified substructures in intermediate or final structures. The peroxide constraint above could be specified on BADLIST, for example. On GOODLIST are placed desired substructures which cannot be entered as superatoms either because their number is uncertain (each GOODLIST entry has an associated minimum and maximum number of occurrences), or because they may share atoms with other superatoms (the "building blocks" must be mutually disjoint

units) or because they are not sufficiently precise: The substructure $X=X-X=X$, where X represents any non-H atom, may be used on GOODLIST to require a conjugated system, but because it is not composed of specific chemical atoms, it cannot be used as a superatom. Two other constraint types, GOODRINGS and BADRINGS, are similarly used to require or forbid the presence of user-specified ring sizes in final structures. The PROTON constraint, especially useful in specifying information from proton NMR spectra, is used to specify desired numbers of protons in specific chemical environments, while the ISOPRENE constraint filters out final structures which do not obey the isoprene rule, an important rule in natural-products chemistry.

The remainder of the work on CONGEN during the year has gone into the human engineering needed to develop an easily used, interactive "front end" to these routines. In this category we include EDITSTRUC, an interactive structure editor, DRAW, a teletype-oriented structure display program and the CONGEN "executive" program which ties the individual pieces together and aids the user with various tasks, such as defining superatoms and substructures, creating lists of constraints and "building blocks" and saving and restoring superatoms, constraints and structures from secondary storage (disc). The resulting system, for which comprehensive user-level documentation has been prepared, is running on the SUMEX computing facility and is available nationwide over the TYMNET and ARPANET networks. Several chemists are currently using CONGEN to assist them in structure identification problems.

Although it is still an unsolved structure, a recent case for which CONGEN was used exemplifies the current capabilities of the program. The compound is a marine sesquiterpene of empirical formula $C_{15}H_{24}$ (with unsaturation, that is number of rings plus multiple bonds, equal to four double-bond equivalents) and there is quite a bit of spectroscopic and chemical data available. These data indicate the presence of a superatom (arbitrarily) called Z, corresponding to structure N+6, along with an isopropyl group (superatom IP) and one



N+6

additional methyl group (superatom ME). These superatoms, together with three more carbon atoms and twelve more protons (or, equivalently, two additional degrees of unsaturation) constitute the "building block" list. There are three parts to

the problem because the superatom Z can have zero, one or two internal bonds created by joining zero, one or two pairs of free valences within Z. Each sub-case represents a separate problem for CONGEN. If no additional constraints are used, the structure generation algorithm constructs 487 intermediate structures for the three cases together, and it can be estimated that the imbedding of Z, IP and ME would produce many thousands of final structures.

For this compound, however, additional information is available. The BADRINGS feature is used to specify that no three-membered rings (which would produce characteristic absorptions in the proton NMR spectrum) may be produced. There is evidence that no multiple bonds other than the exocyclic double bond in Z exist in the molecule, so GOODRINGS is used to require exactly one "two-membered" ring in final structures (CONGEN views multiple bonds as small rings). The fact that there are no methyl groups beyond those in IP and ME is expressed through BADLIST. The final two constraints, that IP must be connected to a methine carbon and ME to a quaternary one, can be entered on either GOODLIST or BADLIST (for the latter, the forbidden environment of IP and ME are specified), and because BADLIST is implemented more efficiently, it is preferred. With these constraints, the structure generation algorithm obtains only 16 intermediate structures for the three sub-cases, in roughly one-third the time required for the unconstrained generation. The imbedding of Z, IP and ME, using the same constraints, give 179 final structures in all.

There is also off-resonance decoupled C-13 NMR data indicating that the compound possesses three methyl groups, five each of methylene and methine carbons and two quaternary carbons. One would expect this data to substantially limit the final structures, but in fact none are eliminated by this GOODLIST constraint. Thus, the "degree sequence" of the carbons is implied by the other constraints, a rather unexpected result.

Additional work on this problem has been carried out using GOODLIST constraints based on analogy with other compounds isolated from the same source and identified. We currently have a set of four structures which represent the most plausible candidates, but a rigorous identification of the compound has not yet been made.

The CONGEN program has just reached the borderline of practical "real world" problems a chemist is likely to face. Although there are some practical cases in which CONGEN has been and is being used, the probability is high that a typical new case, as it is naturally input by the chemist, will either run for an excessively long period of time or will use up all available core storage, or both. There are two facets to this problem. On one hand, the programming has been done primarily in INTERLISP, a language in which the development of complex programs can take place with relative ease. Though this language has assisted in the rapid progress of CONGEN, it is quite inefficient in both execution time and core requirements. We

estimate that the recoding of the most time-consuming portions of the program would speed these portions by a factor of 50-100 and would significantly reduce the demand upon computer memory. On the other hand, there are many cases in which structures are not tested against some of the constraints until after the structures have been generated. In highly constrained cases, this post-testing can be time-consuming; large numbers of structures are generated only to be discarded later. A great deal of research remains to be done in the "intelligent" use of constraints within the program so it can distinguish, early in the analysis, those logical sub-cases which will produce no acceptable structures.

During the next year we plan to address these two problems directly. CONGEN already contains portions in SAIL and FORTRAN, languages much more efficient than INTERLISP, and we plan to recode some of the more time-consuming and less developmental portions of CONGEN into one or the other of these languages. Research, ranging from the discovery of new programming "tricks" to the improvement of the basic mathematics underlying CONGEN, will proceed in the direction of efficient utilization of constraints. Paralleling this will be the continuing development of the "visible" portion CONGEN (the "executive") which, with the guidance of chemist collaborators, will be made increasingly flexible and easy to use. All of these developments will be guided by our desire to make CONGEN a responsive and practical resource for the chemist.

2.3 PLANNER

The DENDRAL PLANNER program [28,33] is designed to analyze the mass spectrum of a compound or of a mixture of related compounds. Because there is no ab initio way of relating a mass spectrum of a complex organic molecule to the structure of that molecule, PLANNER requires fragmentation rules for the class of compounds to which the unknown belongs. This is its major limitation.

Applications and limitations of PLANNER have been discussed extensively. [28,53] The program is very powerful in instances where mass spectrometry rules are strong (i.e., general, with few exceptions). In instances where rules are weak or nonexistent, additional work on known structures and spectra may yield useful rules to make PLANNER applicable (see INTSUM and RULEGEN, below). One unique feature of PLANNER is its ability to analyze the spectra of mixtures in a systematic and thorough way. Thus, it can be applied to spectra obtained as mixtures when GC/MS data are unavailable or impossible to obtain.

The power of the PLANNER has been substantially increased by including the MOLION program (discussed below) as a subroutine for computing the list of plausible molecular ions. Since this subprogram does not depend on knowledge of the compound class, the PLANNER no longer needs to have class-specific rules for determining the mass and empirical formula of the unknown molecule.

The major developments to the PLANNER program have been in the so-called "user interface" - the language and prompts typed by the program to the chemist user. Once the program was successfully transferred to the SUMEX machine, including rewriting parts of it, it was possible to make the program truly interactive. It requires three pieces of information as input from the chemist: the high or low resolution mass spectrum, the characteristic skeletal structure for molecules in the specific compound class, and the fragmentation rules for the class. Additional knowledge about the unknown can be used optionally by the program.

The interactive nature of the program has been enhanced by development of "help" facilities. For example, when a person does not know how to respond to one of the program's prompts, he can always get some help by typing a question mark. An annotated typescript of an interactive session of the PLANNER is shown in Appendix 1.

A chemist's interaction with this program has also been simplified by providing the structure definition and drawing facilities of the CONGEN program (discussed above). Thus it is easy now to tell the program the skeletal structure of the molecules in the compound class. The whole package is relatively natural and easy for a chemist.

We have also added save and restart capabilities to this program so that a chemist can avoid redefining the parameters for a class. This is useful when a chemist needs to explain spectra from more than one sample from the same class. This capability also allows cumulation of knowledge of different classes. For this purpose it is still primitive, but it is a necessary first step.

Along with making the program's parameters accessible to chemist-users, we have also been making the parameters more general to increase their power. For example, the parameter CONTROLRULES, which controls the way the program thresholds evidence, can be set by a chemist to apply to specific individual fragmentations (instead of all fragmentations uniformly). This is a useful way of communicating to the program the heuristic that a given fragmentation process is strong enough and reliable enough that only the strongest evidence for this fragmentation needs to be considered at first. (The program can also be instructed to relax this restriction on this fragmentation later.)

2.3.1 Future Plans

Since the PLANNER program has tentatively moved from a research program to a working laboratory tool, we have discovered a number of ways to tailor the program to the needs of mass spectrometrists. These fall into two general categories: making the program easier to use and making the program more powerful.

The interface will be made "smoother" so that chemists will have an easier time with the interactive dialogue. This means both improvements to the language of the dialogue and improvements to the control structure. The latter is necessary to help the program recover from the errors and from user interruptions.

Additional heuristics will be put at the disposal of users. In particular, specific structural features may be thought to be present or absent in the unknown and we want the PLANNER to use this information as easily as the CONGEN program does. Another improvement will be coupling this program with CONGEN.

We are also ready now to increase the scope of the program by making it work with small structural skeletons and fragmentations involving substituents on a skeleton. For example, aromatic acids are probably best described as an aromatic ring with various substituents. Since the aromatic ring itself does not fragment in characteristic ways, all of the breaks must be described as breaks in the substituents. These improvements mean that many more classes of molecules will be amenable to analysis and many extra types of fragmentations can be used in the analysis.

Finally, we are planning to increase the efficiency of the program. We have already made some of the preliminary timing tests so we know where to focus our attention first. Secondly, we can also reduce system overhead by compiling the program in blocks, which we plan to do.

2.4 CLEANUP

The raw data obtained from GC/LRMS analysis of complex mixtures (e.g. urine samples) consists of a large number (600 to 1000) of mass spectra that result from sampling the GC effluent over an extended fractionation period. Because of the limited separation capability of the GC and contamination by the liquid phase of the GC column most of the spectra obtained are not directly useable. We have developed a program called CLEANUP which takes these raw mass spectral data and removes column bleed and contributions from partially overlapping neighboring elutants.

The CLEANUP program outputs a set of "clean" mass spectra suitable for library matching and/or analysis by other DENDRAL programs. The data reduction factor from the raw data to the cleaned up data is usually an order of magnitude depending on the complexity of the mixture.

The CLEANUP program is directly linked to our MOLION and library search programs. This makes it possible for us to go through the process of data-collection, data-reduction and library search in a smooth automated mode. We obtain as output from this process a line printer listing of possible compounds

for each component found in the mixture. Steps are being taken to extend the system so that components found that are not in the library will be automatically flagged for later reference and subsequent analysis by other DENDRAL programs. We are in the process of writing a manuscript which will describe the procedure of CLEANUP in detail, with examples.

2.5 MOLION

After running a mixture through the GC/MS and CLEANUP, we are left with a collection of more or less pure spectra of unknown compounds. Structure elucidation now begins in earnest. The key elements in problems of structure elucidation are the molecular weight and formula of a compound. Without these absolutely essential data, the structural possibilities are usually too immense to proceed further. Mass spectrometry is frequently used to determine molecular weights and formulae. However, there is no guarantee that the mass spectrum of a compound displays an ion corresponding to the intact molecule. For example, many of the amino acid derivatives in urine samples display no molecular ion. When we are given only the mass spectrum (and for this type of GC/MS analysis a mass spectrum may be all that is available) we must somehow predict likely molecular ion candidates. The new program MOLION [45] performs this task. Given a mass spectrum, it predicts and ranks likely molecular ion candidates independent of the presence or absence of an ion in the spectrum corresponding to the intact molecule.

The MOLION program is written to operate on either low or high resolution mass spectra. It is insensitive to the type of compound analyzed. The program has certain limitations which have been summarized in detail previously.[45] Briefly, the program has difficulties with spectra containing ions from higher molecular weight impurities (thus, the need for CLEANUP, discussed below) and with spectra which contain only low mass ions (representing less than half the weight of the intact molecule). These, of course, are instances where manual interpretation also has most difficulties. The program is currently being modified so that it can cope with spectra of mixtures of compounds.

MOLION is available on SUMEX. A FORTRAN version, initially for low resolution mass spectra, is being written so that the program can be run on smaller computers and exported to others. However, it will continue to be available via SUMEX so that others can access it easily. It is available as a stand-alone, interactive program and also as a part of the PLANNER program.

2.6 Library Search

Over the course of several years, libraries of mass spectral data have been assembled (e.g. GE-network library developed at NIH, the Markey library, the Aldermaston library).

We are presently using spectra from these libraries together with our own compilations to augment our library search facility. Library search provides us with an efficient mechanism for weeding out from a group of spectra those which represent known compounds. Known, commonly occurring components, usually make up more than half of any typical biological sample that we analyze. Clearly, one should spend time on solving the structures of unknown compounds, not rediscovering old ones. The CLEANUP program provides mass spectra which are of sufficient quality to expect that known spectra should be identified relatively easily from such libraries. (Without representative spectra, only total nonsense will result from matching spectra to a library.)

Experience with available libraries has shown that even though these compilations are extensive they are not entirely adequate for analysis of spectra from urine extracts. To cope with this inadequacy we have made an effort to make our library management facility as flexible as possible for updating and modification. As new compounds are identified they are added to our own library compilations for future use.

We are currently investigating the possibility of using a low resolution version of MOLION to enhance the selectivity of the current algorithm. GC retention indices are assigned to the spectra we collect. Some use may be made of these indices in future versions of the program should they be needed.

2.7 MetaDENDRAL Rule Formation Programs

The INTSUM program [34] is in routine, production use to assist in interpretation of the mass spectra of new classes of molecules (see Part 3 for details). When the mass spectrometry rules for a given class of compounds are not known, the INTSUM, RULEGEN and RULEMOD programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the number of molecules in whose spectra there is evidence for each particular fragmentation, along with the total (and average) ion current associated with the fragmentation.

The RULEGEN program attempts to explain the regularities

found by INTSUM in terms of the underlying structural features around the bonds in question that seem to "drive" the fragmentations. For example, INTSUM will notice significant fragmentation of the two different bonds alpha to the carbonyl group in aliphatic ketones. It is left to RULEGEN to discover that these are both instances of the same fundamental alpha-cleavage process that can be predicted any time a bond is alpha to a carbonyl group.

A new development has been the RULEMOD program for modifying and condensing the set of rules produced by INTSUM and RULEGEN together. It looks at the negative evidence associated with each candidate rule in order to select the best ones, then merges rules that seem to explain the same breaks (if possible).

The Meta-DENDRAL programs RULEGEN and RULEMOD have now developed to a point that the programs have rediscovered the mass spectrometry rules for two classes of chemical compounds and substantially aided in the search for explanatory rules for a new family. Chemists are now able to use preliminary versions of these programs profitably. We have used eleven aliphatic amines (and their low resolution spectra) and ten estrogenic steroids (and their high-resolution spectra) as test cases. And we have just run thirteen keto-androstanes as the first test of the new programs on a previously-uncharacterized class of molecules. Although we are attempting to find characteristic rules without necessarily finding explanations for all the data, the final sets of rules are sufficient to explain between one-third and two-thirds of the total data (measured as total ion current).

2.8 Results

2.8.1 Amines

The low resolution spectra of the aliphatic amines are highly ambiguous insofar as many different processes can be generated to explain any one peak. In INTSUM, therefore, we limited the range of interesting processes to those producing a nitrogen-containing ions (high resolution mass spectra of some of these amines reveal that almost all ions contain the nitrogen atom). Processes breaking one or two bonds (with +2 to -2 hydrogen transfers) were considered. The output from INTSUM correlates peaks with breaks that are n bonds away from the nitrogen atom. The output is voluminous because there is some low resolution peak corresponding to almost every one- and two-bond fragmentation for these amines. Also the INTSUM output shows almost no consistent regularities because it looks for regularities in terms of a fixed skeleton which, in this case, is small (-N-).

RULEGEN reduces the break information to eleven rules -- a rule mentioning a bond environment (a subgraph) and a set of bonds in that environment that can be expected to break, with

hydrogen transfers. RULEGEN is attempting to explain the regularities noticed by INTSUM in terms of the bond environments that "drive" the common processes. Each rule is well-supported by the data, but there is still some redundancy in the rules. That is, RULEGEN selects individual rules on the basis of evidential strength without considering the set of rules as a whole.

A new program, RULEMOD, reduces the output of RULEGEN still farther by selecting the "best", or most comprehensive, of the rules. The guiding principle here is that a mass spectral peak only needs to be explained once. (There are cases where this is known to be false -- peaks get contributions from several different processes -- but because there is no way to say in advance which peaks deserve more than one explanation, we let economy of rules guide the selection.) RULEMOD first selects the rule that explains the most peaks (1), then removes those peaks from the evidence supporting the other rules. Then, recursively, the program selects the next best, and so on until the remaining rules have no evidence that is not already explained. For the amines, this step reduces the eleven rules to five.

RULEMOD then considers the possibility of refining the rule set still more by "merging" rules that are very similar. If the subgraphs defined in rules R1 and R2 differ by very little -- i.e., almost every time R1 applies R2 will apply, and vice versa -- then RULEMOD looks for a slightly more general form of the subgraph that will include both R1 and R2 and that will not generate any new negative evidence. This is the first point at which negative instances are considered because, up to here, the number of rules is too large to graph-match against all molecules. For amines, this refinement step rejected all the mergings considered. So the final set of rules for the aliphatic amines, explaining two-thirds of the total ionization of all the spectra, were the five selected previously: alpha cleavage and four two-bond fragmentations (alpha cleavage + cleavage next to N; alpha cleavage + beta cleavage; beta cleavage + cleavage next to N; breaking off two alkyl fragments to keep an arbitrarily large nitrogen-containing fragment).

2.8.2 Estrogens

The high resolution spectra for estrogens made the INTSUM output more specific than for amines, but it was still voluminous. Evidence was gathered for all processes in which either 2 or 3 bonds were broken in the estrogen skeleton (without breaking the aromatic ring, without breaking two bonds to the same carbon, etc.). In each molecule four to ten fragmentation processes showed strong evidence in the corresponding spectrum.

(1) "most" can be defined with respect to number of peaks, percent ionization, or almost any other measure -- we used number of peaks here, giving a higher weight to peaks that can be explained only by the rule in question.

RULEGEN produced 26 rules explaining the INTSUM data (INTSUM looks for regularities and does not try to "cover" all of the data points). RULEMOD selected 15 of those 26 that could explain all of the peaks explained by the 26 rules. Then, in the merging step, RULEMOD merged five pairs of rules together to produce 10 rules explaining one-third of the total ionization in the estrogen spectra. These included descriptions of the well-known breaks B,C,E,F (but not D), cleavage of ring-B next to the ring B-C fusion, cleavage of ring-C next to the ring C-D fusion, and four other processes involving three bonds each. These rules would be considered "good" (generally useful) rules by a mass spectroscopist. Thus, the step of interpreting the INTSUM output and removing ambiguities is now amenable to automation.

2.8.3 Keto-androstanes

The keto-androstanes have not been previously studied and are not as well-behaved as the estrogens. The INTSUM output from the high resolution spectra shows no consistent regularities involving one or two bonds in the environment of the keto group. There is evidence for many fragmentations, but none of it overwhelmingly recommends any fragmentation as being universal or even common among all the members of the class.

RULEGEN first looked at subgraphs large enough to encompass alpha-cleavage (up to 2 atoms away from the bonds broken), but still found no regularities for the whole set of data. We enlarged the bond environment to encompass beta cleavage (up to 3 atoms away from the bonds broken) and tried again. Over 80 rules were produced as possible explanations of the one and two-step fragmentations noticed by INTSUM. About one-third of the total data (36%) were explained by these rules. There was considerable overlap in the rules, because RULEGEN's local view of any rule prevents it from noticing that very similar descriptions of bond environments were produced at different points in its search. RULEMOD found that 20 of these rules could explain all of the data that the 82 rules explained. None of the proposed mergings were acceptable to RULEMOD because it allows a result of merging to have no additional negative evidence that the merged rules alone did not have. (This is a very conservative strategy and needs more study.) This set of 20 rules could be reduced to 13 rules by throwing away the 7 rules whose evaluation scores are below zero (indicating that these rules had more negative evidence than positive evidence). This would reduce the total amount of data explained from 36% to 27% but increases the human readability of the rule set. Such thresholding will be added to the program as an option. We are currently examining additional keto-androstanes and will describe this work shortly.

2.8.4 Future work

2.8.5 New Experiments

The INTSUM, RULEGEN, RULEMOD sequence works well for characterizing a set of molecules. If the set is representative of the whole class then the rules are general, otherwise they may be severely limited in scope. We want to give the program a sense of its limitations so that we can impart it to the chemists.

In particular, we want the program to request new data that will, in some sense, put the inferred rules to the test. And we want to be able to interpret the results of those experiments as indications that particular rules are weak, overly general, etc.

2.8.6 Refinement of Parameters

We want to refine the parameters controlling the operation of the program so that there is a good "default" mode of operation. Because different persons like to focus on different aspects of problems, we want to leave open the option of a person setting the parameters for a very specific purpose. In order to do that, we need to characterize them in a way that chemists readily understand them.

2.8.7 Interactive Programs

Part of the above concerns giving the chemist immediate control of the program. INTSUM and RULEGEN are very interactive now. In order to increase the use of RULEMOD (so we can get the feedback we need as well as to make the programs useful) we are about to start on making it interactive and helpful. Among other things we want to be able to answer questions on-line rather than having to look through hard-copy records of the programs' results.

2.8.8 Merging Two Sets of Rules

The RULEMOD program can merge rules from the output of a single run of RULEGEN. We want to extend this capability so that we will be able to merge different sets of RULEGEN results, first from very similar molecules then from different classes of molecules.

3 PART 3: APPLICATIONS TO BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

3.1 Introduction

In our grant proposal we discussed the application of the instrumentation and computer programs described above to the

study of molecular structure problems in a variety of biomedical applications areas. This is our primary research area, and we discussed specific classes of problems and compounds for investigation. We also made it quite clear that our facilities would be made available to wider community of collaborators/users as our resources permitted. Both categories of application, i.e., within our own group, and with an outside group, are described in some detail below.

We have taken several steps toward encouraging a broad community of potential users to call on our facilities. For example, Appendix 2 contains the memorandum which was sent to local persons who had indicated their potential need for our facilities as described in our proposal. A questionnaire sent to members of the American Society for Mass Spectrometry, Committee III on Computer Applications, resulted in about 55 persons (Appendix 3) indicating a desire to know more about access to our programs. Appendix 4 contains the note which was sent to these persons. The same note has been sent to several other persons whom we know from personal contacts might be interested. Because of the nature of their investigations, many of these people receive NIH support.

The availability of SUMEX as a mechanism for resource sharing has made it possible for us to extend access to our programs to a number of people. Without SUMEX, this access would be impossible, and most of our programs (those which are not easily exportable) could be used only by ourselves.

We have coupled the above efforts with publications (see references 45-49) mentioning explicitly the availability of our programs. Through these efforts, together with talks and informal discussions we are slowly building a local and remote user community.

3.2 Applications by Professor Djerassi's Research Group

Our existing grants, outlined below, mesh well with our instrumentation and program development under the present award. Under NIH Grant GM06840 we have been studying natural products from marine sources with major emphasis on terpenoids and sterols. For this work we have been dependent on the use of our 711 instrument for high resolution mass spectrometry which we require for the identification of all new compounds, many of which are present in only very small quantities. We are particularly anxious to have access to GC coupled with a high resolution mass spectrometer because we hope to be able to screen large numbers of marine animals for their sterol content using this technique. In fact one of Prof. Djerassi's graduate students, Mr. R. Carlson, is currently working on a computer program which will automatically "reject" (i.e., detect, note, and not consider further) known sterols and point toward the presence of unknown ones which will then be the subject of further chemical work. In the context of terpenoids we have used

the INTSUM program with great effect in the structure elucidation of a group of new sesquiterpenes based on the novel skeleton I. A typical example of the sesquiterpenes is II, and the PLANNER, INTSUM and CONGEN programs coupled with high resolution mass spectrometry has been very helpful in elucidating its structure and that of other oxygenated analogs, and in understanding better the mass spectral behavior of these compounds.

Partly under NIH Grant No. GM06840 and partly under Grant No. AM04257 we have been working on elucidating the course of the mass spectrometric fragmentation of steroids and terpenoids through the use of labeled analogs. This information is of fundamental importance in order to apply it subsequently to the structure elucidation of new compounds, and here we have frequently needed access to the 711 instrument because of its "superhigh resolution" capability which was needed for distinguishing between compounds containing only C and H and compounds containing C, H and D. Furthermore, we have needed this instrument for metastable defocusing work and in fact hope very much that this will be eventually automated since this would represent a major simplification for much of our mass spectrometric research work.

Our current and proposed work with our programs for computer-assisted structure elucidation is discussed below under headings consisting of program names, which correspond to the programs discussed in Part 2. Much of the effort in application of a program(s) to the mass spectral data implicitly assumes that the data are available. In fact, without the current and future instrumentation effort discussed in Part 1, these program applications would not be feasible.

3.2.1 CLEANUP

The spectral cleanup program, written for ourselves and our collaborators in the Dept. of Genetics, Stanford Hospital (see Local/Stanford Community, below) will be included as part of the GC/LRMS system on the MAT-711 spectrometer. The essential nature of this program to treatment of the data prior to any more detailed examination was discussed previously. Although it is currently being tested, and now used routinely, for LRMS data from a quadrupole mass spectrometer, it is insensitive to the source of these data.

3.2.2 MOLION

This program is currently in routine use as an adjunct to LRMS data analysis subsequent to CLEANUP. It is incorporated as part of PLANNER as one way to detect molecular ions prior to analysis of the spectrum in terms of structure. Like CLEANUP, this is essentially a utility program, but plays a crucial role in applications of the other programs to structure elucidation problems.

3.2.3 PLANNER

PLANNER is currently being used to test the validity and generality of new mass spectrometry rules derived from INTSUM for the compound classes discussed under INTSUM, for example, the keto-androstanes and capnellanes. Such tests are important to ensure that existing rules can be safely extended to new (perhaps unknown) compounds in the same class.

A planned, major application of PLANNER was mentioned briefly above. The screening of marine sterols will use both GC/LRMS and GC/HRMS. PLANNER will be utilized to examine each spectrum and perform the task it does best, deciding where new substituents are most likely to be found about the steroid (cholestane) skeleton. The interactive nature of PLANNER simplifies the task of providing rules of mass spectrometry and constraints to the program. The rules in the case are the known mechanisms of fragmentation of various classes of sterols. In this way, known substances can be readily identified, even in the presence of other components, and suggested structures obtained for unknown compounds.

3.2.4 INTSUM

As a means of extending the rules of fragmentation in mass spectrometry, several classes of compounds are under study as we attempt to determine characteristic modes of fragmentation. The following is a brief description of each such class and the current status of our research:

1. Pregnanes: Pregnanes related to the progesterone skeleton have been analyzed in some detail in collaboration with Dr. S. Hammerum, (University of Copenhagen, Denmark). One manuscript on this work has been accepted for publication in TETRAHEDRON [51]. One manuscript has been submitted to STEROIDS [52].
2. Androstanes: Keto-substituted analogs of the skeleton of the important steroidal hydrocarbon, androstane, are being studied in collaboration with Dr. Roy Gritter (an IBM scientist who spent his sabbatical leave in our laboratory learning more about mass spectrometry). This study is important to our understanding of the mass spectral behavior of complex, polycyclic systems. It is providing a model for the use of RULEGEN (see below). We are currently analyzing existing data and must acquire additional mass spectral data on new compounds.
3. Macrolide Antibiotics: We are in the middle of an interesting investigation of the fragmentation of macrocyclic antibiotics related to methmycin and neo-methmycin. INTSUM is proving very valuable in determining the regularities in the fragmentation behavior of these polyfunctional and polyheteroatomic compounds.
4. Phytoecdysones: These analogs of insect moulting hormones present difficult analytical problems in analysis of their

structures. We have been analyzing the mass spectra of several of these compounds to determine the feasibility of using PLANNER as a means of screening mixtures of these compounds for new structural types. Unfortunately, their tendency to dehydrate under even the most careful experimental conditions within the mass spectrometer means that spectra are obtained which do not contain as much structural information as we would like. We have confirmed, corrected and extended previous worker's results on diagnostic fragment ions and our investigations are continuing.

5. Insect Juvenile Hormones: In collaboration with Dr. Loren Dunham, Zoecon Corp., we are investigating regularities in the fragmentation behavior of the juvenile hormones. Previous work on the mass spectra of these compounds was carried out only at low resolving powers. So the simple determination of the HRMS will allow re-examination of past results. We have currently a set of representative compounds and their spectra. INTSUM work is now beginning.

3.2.5 RULEGEN

As described above, RULEGEN can be used to assist in discovery of mass spectrometry fragmentation rules which depend on substructural features of molecules. Thus, it can be used for classes of compounds where the fragmentation does not depend on the basic skeleton, but on the positions of substitution. The keto-androstanes represent a case in point, as the fragmentations of these compounds are complex functions of the position of the keto group and the androstane skeleton itself. We are currently using RULEGEN on both "simple" classes of compounds and the estrogenic steroids as well as the androstanes.

3.2.6 CONGEN

We are currently engaged in efforts to explore the utility of CONGEN to a variety of structure elucidation problems. The current areas of application are summarized below, together with progress to date.

- 1) Chlorinated Hydrocarbons: As an important class of a more general problem area in chemical structure analysis, the isomerism various types of chlorinated hydrocarbons has been investigated. The general structural problem is identification of possible substitution isomers about a given skeleton. The chlorinated hydrocarbons are, in addition, an important environmental and health problem. Using the labelling algorithm [41] we have constructed possible isomers of several classes of chloro-carbons [46].
- 2) Vertex-Graphs and Ring Systems: CONGEN has been used to explore possible ring systems in organic chemistry [47]. These features of the program have been utilized in several of the studies described below.

- 3) Ion Structures: CONGEN has been used to construct possible ion structures under a variety of constraints in support of studies on the structures of ions in the mass spectrometer. These studies are crucial to a deeper understanding of molecular fragmentation. The programs results (manuscript in preparation) are used to ensure that no plausible alternatives have been overlooked during efforts to characterize the structures.
- 4) Terpenoid Systems: We are using CONGEN to explore questions of the scope of terpenoid isomerism. We would like to determine some criteria which might allow us to say something about why only certain structural types are found in nature, to the exclusion of many possibilities which are very similar in structure.
- 5) General Structure Elucidation Problems: We are currently using CONGEN in two modes in connection with the structures of new compounds which have arisen in our recent research on marine organisms. The first mode has been to test several cases for which a structure had been proposed, to ensure that no other reasonable candidates had been overlooked. The second mode is in suggestion of structural possibilities for as yet unknown compounds, as we attempt to narrow the problem further by examination of candidate structures and design of experiments to differentiate among them.
- 6) Scope of Structural Isomerism: We are investigating the philosophical and pedagogical aspects of the scope of structural isomerism. This investigation is important to our program design and strategy as we identify the ways persons consider and reject whole categories of structural possibilities. The important artificial intelligence aspects of CONGEN lie in its ability to reason about molecular structure using efficient problem-solving strategies.

3.3 Applications by Other Members of the Stanford Chemistry Dept.

- 1) Prof. Mosher: We have used CONGEN to suggest structural possibilities for a naturally occurring analog of the fish poison tetrodotoxin. This structure is still under investigation.
- 2) Prof. Hahn (on sabbatical leave at Stanford from Syracuse University): We have used CONGEN to explore possible structures for unknown products of a photochemical reaction. These results have led him to begin a new set of experiments (specifically, CMR) to greatly restrict the possibilities.
- 3) Prof. Johnson: In his wide-ranging syntheses of steroid hormones and other steroids of biological interest, he has studied reactions involving stereo-specific cyclization. We are investigating use of CONGEN for structural analysis under

constraints imposed by synthetic cyclization experiments. For example, a previously investigated compound was found to have two structural possibilities. The new possibility could not be differentiated from the assigned structure based on available data.

- 4) Prof. Collman: We have utilized our mass spectrometry facilities to analyze samples in support of his work on oxygen binding to porphyrins (hemoglobin models).
- 5) Prof. Van Tamelen: We have provided mass spectrometry support (HRMS) to assist in the characterization of several compounds related to his work on terpenoid cyclizations.

3.4 Applications by Other Stanford University Scientists

- 1) Genetics Research Center (GRC) Stanford Hospital: One of our strongest collaborations because of their requirements for additional automation in data reduction and analysis. Their screening program for metabolites characteristic of diseases of genetic origin uses GC/LRMS as the primary source of data. The CLEANUP and MOLION programs were written at least in part to assist the GRC in more systematic approaches to their data. We are currently using CONGEN to assist in determination of structures of unknowns for which mass spectrometric and chemical data are available. Our GC/HRMS facilities will also be utilized for problems which require determination of empirical formulas for ions in spectra of unknown compounds.
- 2) Stanford Pharmacy: We have had several requests for assistance from the Pharmacy of Stanford Hospital (Director: Dr. Hiram Serra). These have variously involved analyzing the stability and purity of pharmaceutical preparations, in particular: a. the impurity of stock preparations b. the stability of nitroglycerine tablets to heat; c. the stability over several months of methyl-dopa, prednisone and banthine when these compounds were formulated into syrups.
- 3) Drug Assay Laboratory Department of Pharmacology, Stanford University: Research personnel from this laboratory (Director: Sumner M. Kalman) have requested mass spectra on various derivatives of digoxin using both high and low resolution data.
- 4) Department of Psychiatry, Stanford University: The research group headed by Dr. J. Barchas has used low resolution mass spectral data for the purpose of structure elucidation of a basic compound of interest to their research program.
- 5) Department of Anesthesia, Stanford University: The DENDRAL group was asked by Dr. J. Trudell to help him in the identification of a urinary metabolite isolated after the administration of an anesthetic. This work involved high

resolution mass spectrometry of fractions isolated by Dr. Trudell.

- 6) Department of Psychiatry, Palo Alto Veterans Hospital: In this work we analyzed samples by GC/MS given to us by Dr. S. Kanter who works with Dr. Hollister. They were interested in detecting cannabinal, delta-9-tetrahydrocannabinol and an unknown (molecular weight 312) from urine extracts of subjects who had smoked marijuana. This involved running standards of cannabinal and its delta-9-tetrahydro analog through the GC/MS. We were unable to identify these compounds by mass spectrometry as being present in urine. In a subsequent meeting we learned that their concentration was less than 20 nanogram (per GC/MS injection) which is below the limits of sample flow for the recording of reproducible mass spectra. Dr. Kanter is working on the problem of isolating sufficient material for GC/MS and we expect to continue this project in year II of the current grant.
- 7) Prof. McCarty - Civil Engineering: Prof. McCarty is involved in a project to monitor water quality of effluents from tertiary sewage plants. This project includes significant efforts at characterization of the organic content of the water in various phases of its treatment to determine the efficiency of removal of various materials and to identify unknown organic compounds. We have agreed to provide instrumental and computer program support where necessary to assist him in characterization of these samples.

3.5 Applications by Non-Stanford Scientists

As an additional component of the resource sharing aspects of research, we have, as resources allow, extended the use of our facilities to a group of users remote from the local Stanford community. We have divided these users into two categories, those for whom we have provided mass spectrometry support and those who represent users of DENDRAL programs and collaborators on program development via the SUMEX resources.

A. Users of Mass Spectrometry Facilities

- 1) Professor O. O. Orazi, La Plata, Argentina: During the past year we have supplied Dr. Orazi with three low resolution mass spectra. We will be providing HRMS data for him in year II of our grant.
- 2) Professor T. Nakano, Caracas, Venezuela: Dr. Nakano sent one sample of an unknown alkaloid for high resolution mass spectrometry. We were able to show that his low resolution mass spectrum was 2 amu from the true molecular ion and after recording a low resolution mass spectrum his alkaloid was identified as a known compound.

- 3) Dr. Steen Hammerum, Copenhagen, Denmark: Dr. Hammerum requested our assistance in running ultra high resolution mass measurements on several ions in the mass spectra of compounds he had specifically labelled with ^{13}C .

B. Users/Collaborators of/with DENDRAL Programs on SUMEX

Below, in alphabetical order, we list those persons who have a) expressed interest in use of our programs and have been sent instructions in how to gain access to SUMEX and our programs. In many cases these persons have received more detailed information in the form of demonstrations in person or remotely using the LINK facilities of SUMEX, and b) persons who have acted as collaborators in development of parts of one or more of our programs. (Some persons fall in both categories.) Each prospective user generally receives the following packet of material:

i) An Introduction to SUMEX-AIM - a simplified guide to the SUMEX system for those who do not need the complete TENEX operating system manual.

ii) Network information - TYMNET or ARPANET instructions on how to connect to the network and use the network to connect to SUMEX.

iii) Account and password information.

iv) Documentation for the specific programs the user wishes to access.

v) Explanation of the SUMEX RECORD facility which allows us to examine a user's complete terminal session when he has encountered trouble.

Because we have just begun encouraging a significant community of persons to try our programs, we do not yet have a good idea of which persons will continue as serious users. But we have at least provided the opportunity for persons to gain access to our programs, try them and determine how they might (or might not) fit into their own research problems. The term "exploratory" refers precisely to this category of persons who are now engaged in this kind of evaluation. The program names after each person's activity refer to their current major interest. In some cases, we do not actually know the specific problems which are being explored.

1. Dr. A.L. Burlingame (U.C. Berkeley) - Exploratory - all DENDRAL programs.
2. Prof. E.J. Corey (Harvard) - Exploratory, collaboration on programming strategies, CONGEN.
3. Dr. L. Dunham (Zoecon) - Exploratory - Structure determination - CONGEN.

4. Dr. H.M. Fales (NIH) - Exploratory - all DENDRAL programs.
5. R. Feldmann (NIH) - Collaborative development of programs (structure input and drawing routines).
6. Prof. D.L. Fishel (Kent State) - Considering access to SUMEX - has the program descriptions.
7. Prof. M.J. Goldstein (Cornell) - We have provided CONGEN results to him for a difficult structure problem.
8. Dr. N.A.B. Gray (Cambridge) - Collaborating on strategies for computer-assisted structure elucidation programs. He is working on spectral data interpretation.
9. Dr. P. Gund (Merck, Sharpe & Dohme) - Arranging an on-line demonstration for exploratory purposes - CONGEN.
10. Dr. J. Karliner (Ciba-Geigy) - Using CONGEN on structure elucidation problems.
11. Dr. S. Heller (Environmental Protection Agency) - Collaboration on mass spectral library development.
12. Dr. P. Jurs (Penn. State) - Collaboration on structure analysis and building of chemical structure models.
13. Dr. B. Kowalski (Univ. of Washington) - Has approached us for use of SUMEX in pattern recognition work.
14. Dr. D. Lefkowitz (Univ. of Penn.) - Exploratory - interest in DRAW portion of CONGEN for NCI chemical information system.
15. Dr. S. Markey (NIH) - Exploratory - all DENDRAL programs.
16. Dr. F. McLafferty (Cornell) - Exploratory - all DENDRAL Programs.
17. Dr. R. Milberg (National Center for Tox. Res.) - Exploratory - CONGEN.
18. Dr. D. Poulter (Univ. of Utah) - Exploratory - use of CONGEN in structure determination problems, especially terpenoids.
19. Dr. K. Rinehart (Univ. of Illinois) - Exploratory - all DENDRAL programs.
20. Dr. P. Roller (National Cancer Institute) - Exploratory - all DENDRAL programs.
21. Dr. R. Rosen (FMC Corp.) - Exploratory - all DENDRAL programs.
22. Dr. G. Szonyi (Polaroid Corp.) - Interest in CONGEN, exploratory phase beginning.

23. Dr. W.T. Wipke (Princeton) - Exploratory - CONGEN collaboration on structure model building and methods for stereochemical representation of chemical structure.

4 BIBLIOGRAPHY

See Section II-D, SUMMARY OF PUBLICATIONS, for a listing of publications by this project.

Figure 1. Mass Spectrometry
Data Acquisition Hardware
Configuration

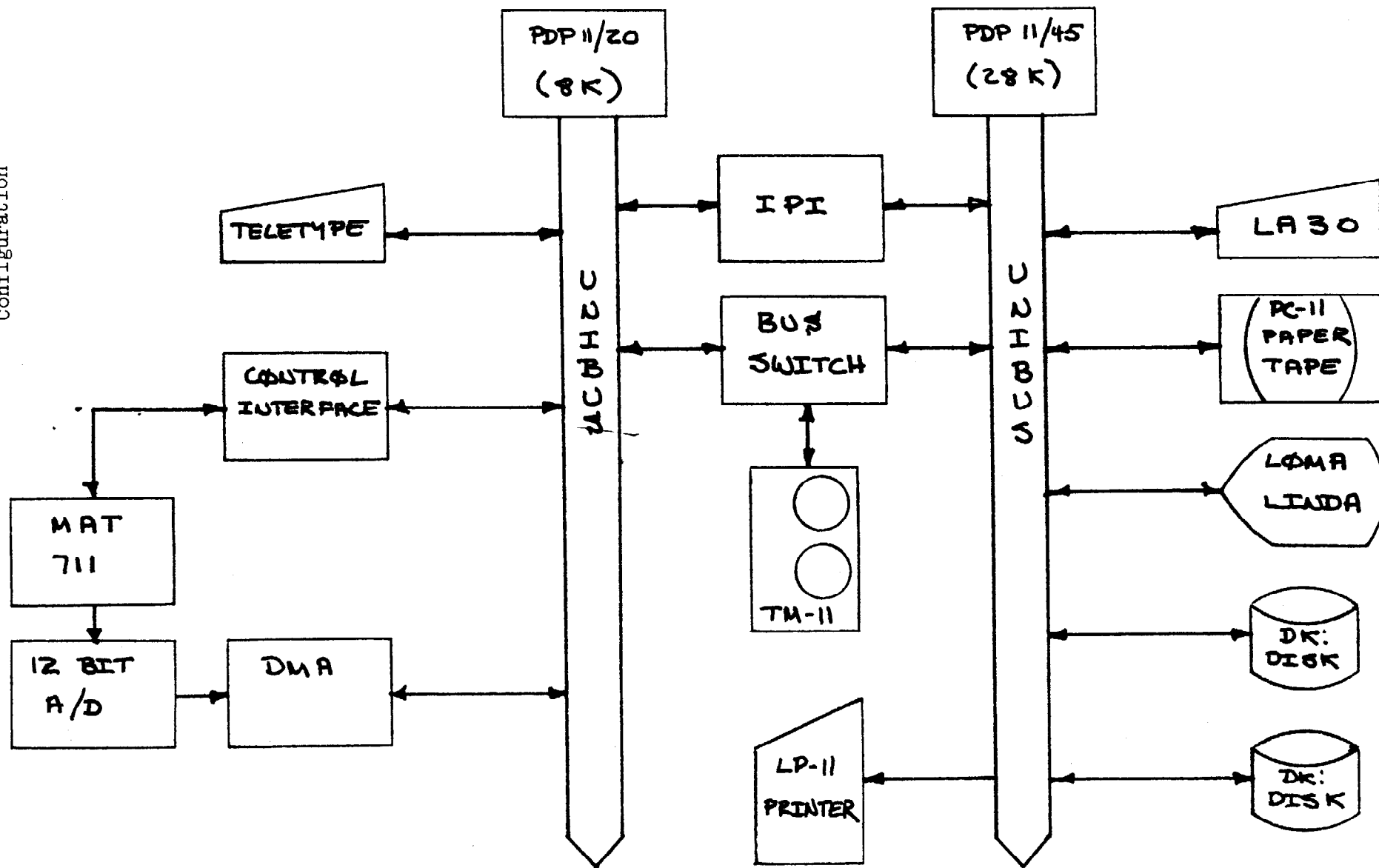
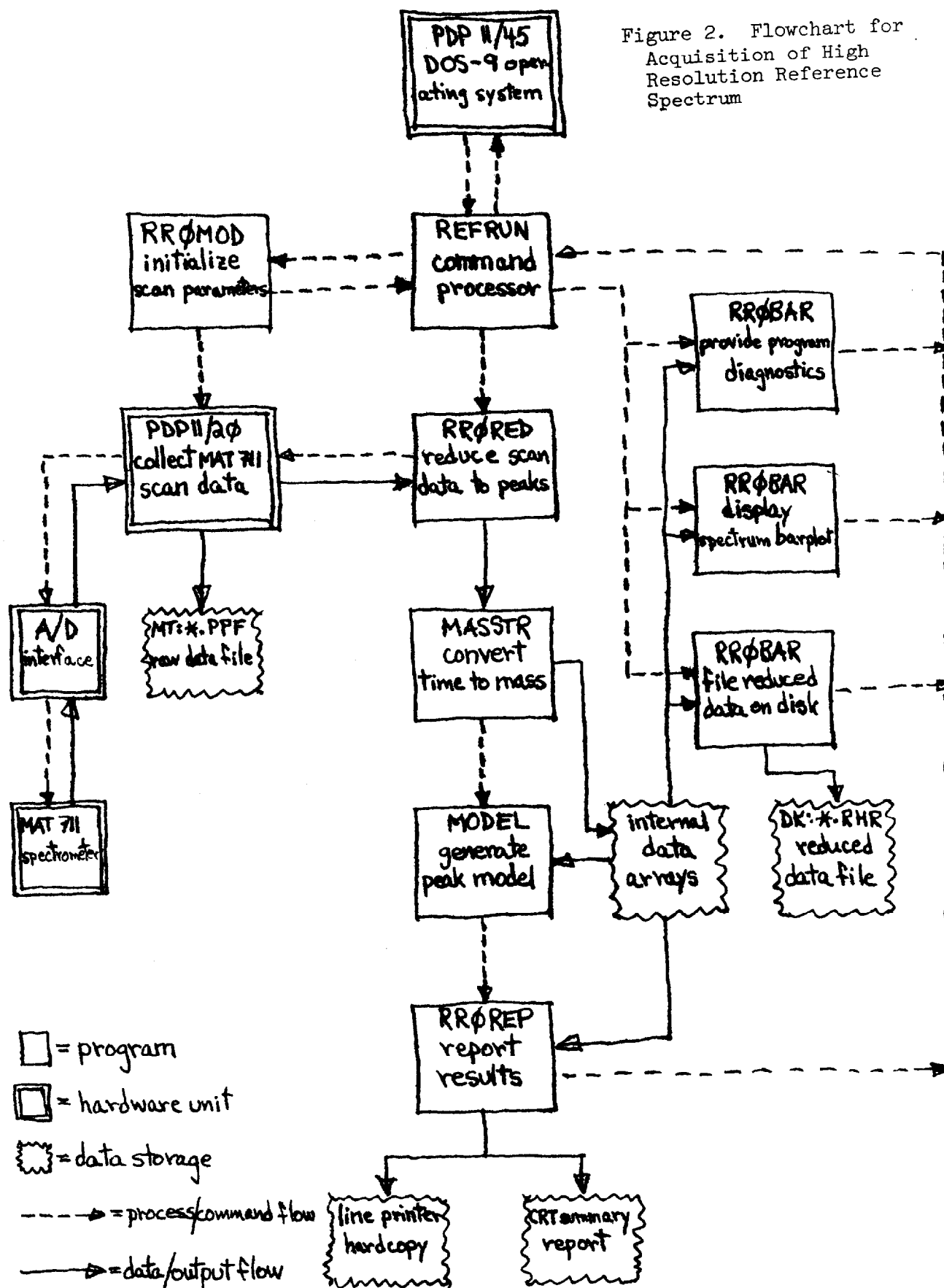
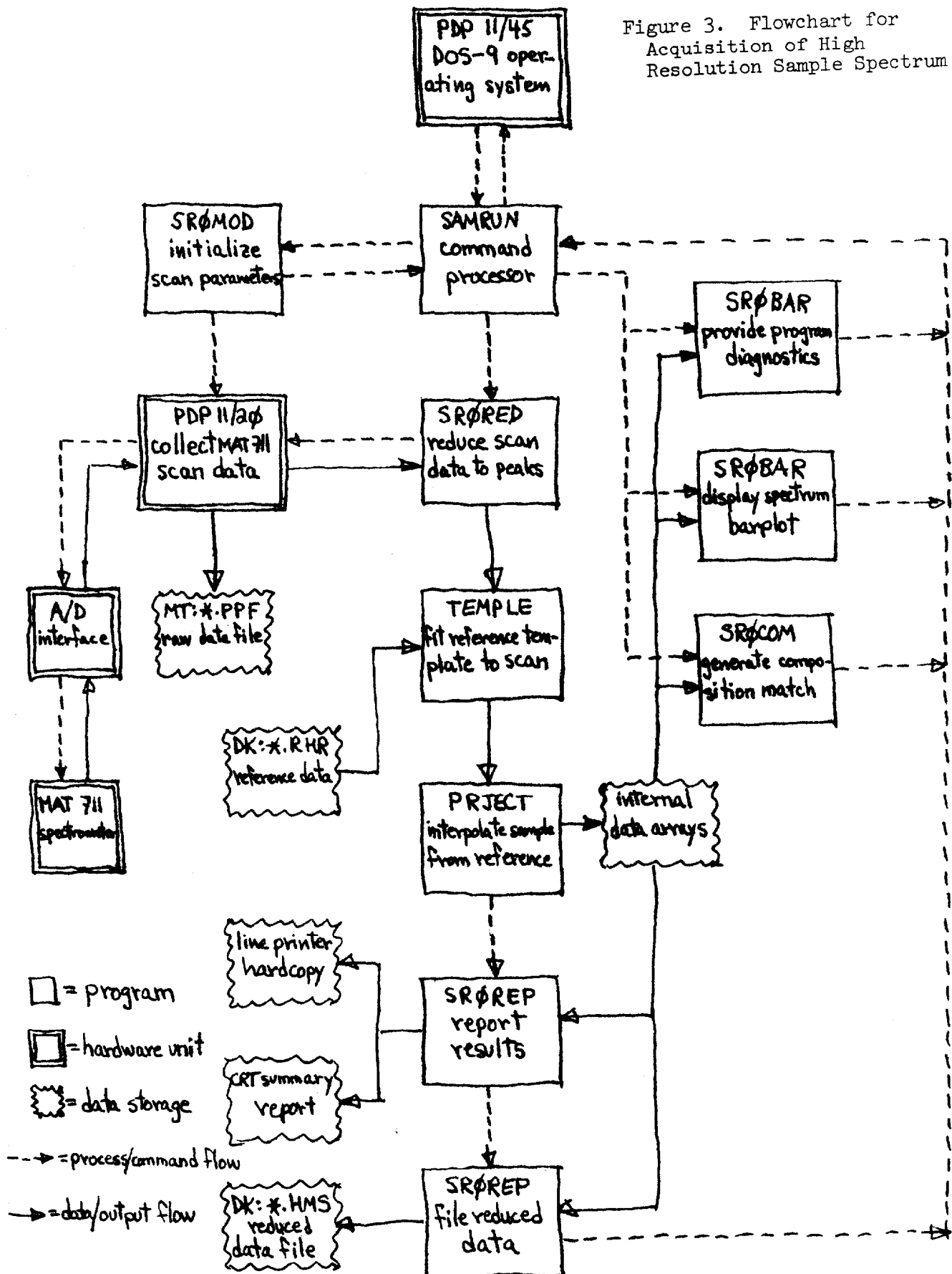


Figure 2. Flowchart for Acquisition of High Resolution Reference Spectrum



Flowchart For Acquisition of High Resolution Reference Spectrum

Figure 3. Flowchart for Acquisition of High Resolution Sample Spectrum



Flowchart for Acquisition of High Resolution Sample Spectrum

Appendix 1
Typescript of Interactive Session with
The DENDRAL PLANNER

Annotations are bracketed and prefixed with asterisks.

[*** 1. Start the Planner program.]

@WORK5

INTERLISP-10 20-OCT-74 ...

Good morning, Bruce.

(WORK5.;1 . <SUBSYS>NLISP.SAV;1)

_RUN]

[*** 2. After the program prompts for class name it asks if the structure and fragmentation rules for the class have been stored in a file from a previous session.]

CLASS NAME:CAPNELLANE

GET CLASS PARAMETERS FROM A FILE (Y/N) ? Y

GET CLASS FILE NAMED: <SMITH>CAPNELLANE.PLANTEST

FILE CREATED 20-MAR-75 14:08:31

&

YOU HAVE JUST SET PARAMETERS FOR THE CLASS CAPNELLANE

SKELETAL COMPOSITION: ((H . 26) (C . 15))

[*** 3. The data are now read in from a file and adjusted for isotopic contributions, with peaks below the noise threshold deleted.]

READ A NEW MASS SPECTRUM (Y/N) ?Y

NAME OF SPECTRUM FILE: <SMITH>1CAPNELLANE.TEST

HIGH RESOLUTION

MASS RANGE: 40.0317 TO 234.1621 (82 PEAKS)

TOTAL ION VALUE: 1440.652

(PEAKS SMALLER THAN 7.203262 DELETED.)

PRINT ? N

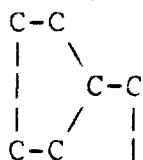
[*** 4. The program is ready to begin its analysis at this point. However, the user asks instead to see the structure and breaks for this class, to be sure that they were set correctly in the previous session.]

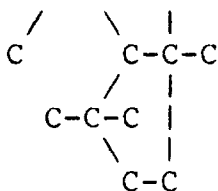
START PLANNING (Y/N)?N

WANT HELP (Y/N)?N

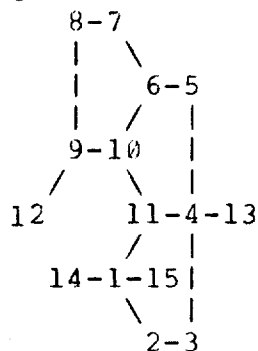
STRUC?

CAPNELLANE





CAPNELLANE



BREAKS?

 LABEL: 6H
 BONDS: ((3 - 2) (1 - 11))
 HTRANS: (-1 0)

LABEL: 7H
 BONDS: ((11 - 1) (3 - 4))
 HTRANS: (-1 0 1 2)

LABEL: 7L
 BONDS: ((1 - 11) (3 - 4))
 HTRANS: (-1)

[*** 5. When the user examines the parameter CONTROLRULES, it has no effective value. So it is given a value in order to control the amount of evidence used by the program to build structures.]

CONTROLRULES?
 CONTROLRULES= NIL
 CONTROLRULES
 USE ONLY THE STRONGEST EVIDENCE FOR SOME BREAKS (Y/N)?Y
 APPLICABLE BREAKS:6H
 APPLY ON PASSES:?

ENTER EITHER A LIST OF INTEGERS OR THE WORD ALL. THE INTEGERS INDICATE THE PASS NUMBERS (I.E. THE NUMBERS OF THE ATTEMPTS) OF THE STRUCTURE BUILDING PGM FOR WHICH THIS CRITERION IS APPLIED TO THESE BREAKS. E.G. PASS 1 BUT NOT AFTERWARD, PASSES 1 AND 2 BUT NOT AFTERWARD, OR ALL PASSES.

PASSES:1
 USE EVIDENCE THRESHOLD FOR SOME BREAKS (Y/N)?Y
 APPLICABLE BREAKS:7H 7L
 APPLY ON PASSES:1
 PERCENT OF MAXIMUM INTENSITY (DEFAULT = 33): ?

WHAT THRESHOLD DO YOU WANT TO USE AS A CUTOFF -- GIVE A NUMBER THAT
WILL BE USED AS PERCENT OF MAXIMUM INTENSITY TO THROW AWAY (RELATIVELY)
SMALL INTENSITY PIECES OF EVIDENCE.
PERCENT: 33

[*** 6. Now we are ready to begin analyzing the mass spectrum.]

PLAN

[*** 6a. The program first infers that the molecule contains
15 carbons, 22 hydrogens, and 2 oxygens.]

(COMPUTING MOLECULAR ION(S))

MOLECULAR IONS

(234.1609 100 (C . 15) (H . 22) (O . 2))

[*** 6b. The program now looks for evidence (peaks in the mass
spectrum) corresponding to possible combinations
of residual atoms and double bonds in each fragment.]

(STARTING ANALYSIS PART)

BREAK : SUBSTITUENTS ON CHARGED FRAGMENT : EVIDENCE (M/E)

6H	((DOT . 4) (C . 0) (O . 1))	161.0975
	((DOT . 2) (C . 0) (O . 1))	163.1136
	((DOT . 4) (C . 0) (O . 2))	178.0993

7H	((DOT . 4) (C . 0))	133.1009	
	((DOT . 2) (C . 0))	133.1009	
	((DOT . 4) (C . 0) (O . 1))	150.1038	149.0951 148.088
147.0811	((DOT . 2) (C . 0) (O . 1))	150.1038	149.0951
	((DOT . 4) (C . 0) (O . 2))	163.0758	

7L	((DOT . 4) (C . 0))	65.03971
	((DOT . 2) (C . 0))	67.05513
	((DOT . 0) (C . 0))	69.07053
	((DOT . 0) (C . 0) (O . 1))	85.06461

[*** 6c. Structural descriptions are now put together in all
plausible ways that are consistent with the substituent
evidence just gathered.]

BEGIN SYNTHESIS OF MOLECULAR ION = 234.1609

STRUCTURE 1

((DOT . 4) (O . 1)) C4 C5 C6 C7 C8 C9 C10 C11 C12 C13)
((O . 1)) C3)

EVIDENCE USED TO BUILD STRUCTURE:

(7H (DOT . 4) (O . 1))
(6H (DOT . 4) (O . 2))
(7L (O . 1))

DONE

Appendix 2
Memorandum to Stanford Scientists

DATE: May 2, 1974

TO : Potential Stanford University Collaborators with the DENDRAL Project

FROM : Carl Djerassi, Professor of Chemistry

SUBJECT: Availability of Facilities

During the preparation (December, 1973) of our grant application entitled "Resource Related Research: Computers in Chemistry" your research group expressed interest in utilizing our facilities (see below) for assistance in solving structure elucidation problems related to health sciences. I now wish to notify you, in my role as P. I. for this grant, that it was funded as of May 1, 1974, for a three year period, substantially as requested.

I want to make a few general comments before describing the facilities which will become available during the course of this grant. Our primary goals, as spelled out in our grant application, deal with bringing state-of-the-art techniques in mass spectrometry and computer science to bear on solving problems of molecular structure. We are not funded to act as a general service facility; we have neither the time nor the personnel to function in this manner. We hope to operate in a collaborative manner with each of you to help decide questions concerning the specific instrumental and computer techniques which can be brought to bear on your problems.

Facilities

A) Mass Spectrometry. Our primary goal is to provide the capability for routine gas chromatography/high resolution mass spectrometry. We will shortly receive a PDP 11/45 computer system for the laboratory which will be programmed to carry out this task. At the present time we can provide gas chromatography/low resolution mass spectrometry and severely restricted (without charge) access to high resolution mass spectra of single compounds as our budget provides only minimal funds for supporting this work pending the completion of the 11/45 data system. These facilities are available at no cost to the user.

B) Computer-Assisted Structure Elucidation. We have available a number of programs designed for automatic analysis of mass spectral data and also for isomer generation and manipulation and display of chemical structures. We are designing and programming interactive systems which will allow users to answer problems concerning the identity of molecular structures based on a variety of spectroscopic data. These interactive systems are under development, but are always available in their present state for tolerant users. These programs run on the SUMEX (Prof. Lederberg, P. I.) PDP-10 at the Medical School. Details of local user access are being worked out, but a significant amount of computer time will be available free of charge for users of these programs.

For the present time, those of you who are interested in making use of these facilities should contact (Med. School) Dr. Alan Duffield (Ext. 7-5788 (temporary, soon to be 7-6389) or (Chemistry Dept.) Dr. Dennis Smith (Ext. 7-3144). If serious bottlenecks occur, either with respect to the mass spectrometry laboratory or SUMEX, I intend to make use of an advisory committee described in our grant application to help rectify the problem. It is my hope that judicious selection of problems along the lines of the collaboration I outlined above will not make this necessary.

CD:ab

cc: Prof. J. Lederberg, Genetics
Prof. E. Feigenbaum, Computer Science



Appendix 3
Questionnaire Sent to American Society for Mass Spectrometry,
Committee III on Computer Applications

DENDRAL LIST

<u>Name</u>	<u>Address</u>		
H.E. Lumpkin	Exxon Res. & Eng. Co. P.O. Box 4255	Baytown, Texas	77520
R.K. Mauldin	Celanese P.O. Box 1414	Charlotte, N.C.	28232
M.F. Zabielski	United Aircraft Res. Labs.	E. Hartford Conn.	06108
T.E. Edwards	Mail Stop ES 32 Marshall Space Flt. Ctr. E228/113 Exp. Station	Huntsville, Ala.	35812
C.N. McEwen	Dupont	Wilmington, Del.	19898
B. Tiffany	Savannah River Labs. Dupont	Aiken, S.C.	29801
J. Capellan	Ames Lab USAEC Iowa State	Ames, Iowa	50010
K.R. Thompson	Eng. Phys. Lab B-357 Dupont	Wilmington, Del.	19898
T. Krick	Biochem. Dept. U. of Minn.	St. Paul, Minn.	55101
J.R. Hass	NIEHS P.O. Box 12233	Research Triangle Park, N.C.	27709
F. Hileman	D. of Chem. U. of Utah	SLC, Utah	84112
M. Hoffman	D. of Chem Kansas State U	Manhattan, KS	66506
W.T. Rainey	P.O. Box Y Bldg. 9735 ORNL	Oakridge, TN	37830
M.E. Fitzgerald	Arco Chemical Co. R + E P.O. Box 85	Glenholden, PA	19036
W.L. Krudop	3316 Darryl Ln.	Modesto, Calif.	95350
B.A. Raby	325 N. Mathilda Ave.	Sunnyvale, CA	94086
W.C. Judd	Gen. Elect. 21800 Tungsten	Cleveland, OH	44117
R.W. King	U. of Florida Dept. of Chemistry	Gainesville, FL	32611
W. Killinger	W.R. Grace Co. Wash. Res. Ctr. 7379 RT32	Columbia, MD	21044
M.C. Manning	Continental Oil Co. R + D	Ponca City, Okla.	74601
J. Dillard	Chem. Dept. VA. Tech	Blacksburg, VA	24061
E.M. Chait	Dupont Instruments 1500 S. Shamrock Ave.	Monrovia, CA	91016

<u>Name</u>	<u>Address</u>		
R.T. Rosen	FMC Corp. P.O. Box 8	Princeton, N.J.	08540
J.C. Orr	Harvard Medical School 45 Shattuck St.	Boston, MA	02115
M. Levenberg	Dept. 482 Abbot Laboratories	North Chicago, Ill.	60064
J.M. McGuire	S.E. Environmental Res. Lab. College Station Rd.	Athens, GA	30601
B.S. Middleditch	Dept. Cell Biology Baylor Coll. Med.	Houston, Texas	77025
G.L. Kearns	The Kearns Group 58 Buckingham Dr.	Stamford, CT	06902
H. Fales	NIH	Bethesda, MO	20014
E. G. Perkins	104 Burnside Lab U of Illinois	Urbana, ILL.	61801
M.W. Karasek	Chem. Dept. U of Waterloo	Waterloo, Ontario	
J.M. Miller	D of Chem Brock University	St. Catharines, Ontario	
I.M. Campbell	RM 605 GSPH U of Pittsburgh, 130 Desto St.	Pittsburgh, PA	15261
K.A. Lincoln	NASA Ames Res. Ctr. Moffett Field	CA	94035
F. Falkner	Drug Metabolism Central Res. Pfizer, Inc.	Eastern Pt. Rd. Groton, Conn	06340
D.V. Bowen	1221 York Ave. New York, NY		10021
R.D. Grigsby	Dept. of Biochem and Biophys. Texas A + MU.	College Station, TX	77843
L.R. Yetter	IBM Corp. P.O. Box 417, 19 Barton Rd.	Apalachin, NY	13732
B.A. Allen Jr.	832 Sycamore Lake Jackson	TX	77566
H.S. Hertz	A105 Chemistry Nat'l. Bur. Stds.	Wash. D.C.	20234
R. J. Ayers	FDA 1560 E. Jefferson	Detroit, Mich.	48207
J.F. Paulson	A.F. Cambridge Res. Labs/LKB Hanscom AFB	Bedford, Mass.	01730
R.J. Weinkam	Dept. Pharm. Chem. U of Calif. S.F.	San Francisco, Calif.	94143
K.L. Rinehart	454 Roger Adams Lab U of Ill.	Urbana, Ill.	61801

<u>Names</u>	<u>Address</u>		
C.A. Stearns	NASA-106-1 21000 Brookpark Rd.	Cleveland, Ohio	44135
G. Eigendorf	D. of Chem U of British Columbia	Vancouver, 8 B.C., Canada	
T.F. Thomas	D. of Chem. U. of Mo, KC	Kansas City, Mo	64110
D. L. Winter	Rm. 228 R + D Continental Oil Co.	Ponco City, OK	74601
D. L. Fishel	D. of Chem. Kent State Univ.	Kent, Ohio	44242
G.W. Wilcox	1600 W. Smile Rd. Ferndale, MI		48220
S. Hammerum	D. of General and Org. Chem. U. of Copenhagen DK-2100	Copenhagen, Denmark	
A.M. Hogg	D. of Chem. University of Alberta	Edmonton Alberta, Canada	T6G2G2
S.A. Wikstrom	Med. Univ. of S.C. Pharmacology	Charleston, S.C.	29401
D. J. Harvey	Pharmacology Dept. South Parks Road	Oxford, OX 13dT	
D.E. Games	Dept. of Chem. Univ. College P.O. Box 78 CF1 1XL United Kingdom	Cardiff	
M.A. Grayson	Dept 221 Bldg. 33 McDonnell Douglas Research Labs	St. Louis, Mo.	63166
C.H. Williams Jr.	UNICAMP/QUIMICA 13.100 Campinas Sp.	Brazil	
★J. Lehman	208 Progress Ave.	Hamilton, Ohio	45013
I. SAKAI	BASIC RESEARCH LAB. TORAY INDUSTRIES INC. 1111 TEBIRO	KAMAKURA Z48 JAPAN	

Appendix 4
Letter Sent to Mass Spectroscopists Responding to Questionnaire

DENDRAL Program Availability on SUMEX

The Stanford University Medical Experimental computer facility (SUMEX) has been established at Stanford with the support of the Biotechnology Resources Branch, National Institutes of Health. Its primary mission is resource sharing, where the resources in this case are complex computer programs applied to health-research problems and the computing facility on which these programs can be used via a nationwide computer network. SUMEX is actively encouraging the development of a collaborative community of users of this facility.

The DENDRAL Project at Stanford is one of the initial collaborative projects on SUMEX. This project is making its programs available to the outside community within the limits of available resources. Because resources are limited, the future may bring a more restrictive policy of access than that presently in force. Until that time, however, the SUMEX policy is to encourage as many qualified and interested persons as possible to access the program on a trial basis, to determine the potential applicability of such programs to ongoing research. The major programs available now are outlined below:

- 1) PLANNER--Infers possible structures of unknown compounds (singly or as mixtures) given a mass spectrum and fragmentation rules of the class of compounds to which the unknown(s) presumably belongs. (See Smith et al., J. Amer. Chem. Soc., 94, 5962 (1972); ibid., 95, 6078 (1973)).
- 2) INTSUM--Given a set of known, related structures and the mass spectrum corresponding to each structure, INTSUM suggests possible fragmentation processes which resulted in the observed ions, and then summarizes the results in terms of processes which are general to the class of structures, and those which are specific to certain members of the class. (See Smith et al., Tetrahedron, 29, 3117 (1973)).
- 3) CONGEN--CONGEN (constrained structure generation) accepts as input known structural features of an unknown molecule (whose elemental composition is known) and produces all structural isomers consistent with these data. The features and constraints are entered in an interactive session with the program and results can be drawn at a terminal or further constraints added based on examination of new data. CONGEN represents our initial version of a program for computer-assisted structure elucidation. The structure generator which underlies CONGEN has been described (See Masinter et al., J. Amer. Chem. Soc., 96, 7702 (1973) and ibid., 7714 (1974)).
- 4) MOLECULAR ION DETERMINATION--Given a (low or high resolution) mass spectrum in which the molecular ion may or may not be present, this program suggests a ranked list of candidate molecular ions. (See G. Dromey, B. G. Buchanan, D. H. Smith, J. Lederberg and C. Djerassi, J. Org. Chem., in press (March 1975)).

For additional information on these programs and access to SUMEX, write to Professor Joshua Lederberg, SUMEX Project, Department of Genetics, Stanford University, Stanford, California 94305, or Professor Carl Djerassi or Dr. Dennis H. Smith, Department of Chemistry, Stanford University, Stanford, California 94305. It will be helpful to indicate the basis of your interest and intended applications although it is understood that trial use is a prerequisite to a considered answer to such a question.

Appendix 5
Draft of Manuscript for American Chemical Society Symposium
on Computer Networking in Chemistry

NETWORKING AND A COLLABORATIVE RESEARCH COMMUNITY: A
CASE STUDY USING THE DENDRAL PROGRAMS.

Raymond E. Carhart*, Suzanne M. Johnson, Dennis H. Smith, Bruce G. Buchanan, R. Geoffrey Dromey, and Joshua Lederberg.

Departments of *Computer Science, Genetics, and Chemistry, Stanford University, Stanford, California, 94305.

Computer Science is one of the newest, but also one of the least "cumulative" of the sciences. Gordon (1) has recently pointed out the upsetting disparity between the number of potentially sharable programs in existence and the number which are easily accessible to a given researcher. Although some mechanisms exist for the systematic exchange of program resources, for example the World List of Crystallographic Computer Programs (2), a great deal of programming effort is duplicated among different research groups with common interests. The reasons for this are understandable: these groups are separated by geography, by incompatibilities in computer facilities, and by a lack of a means to keep abreast of a rapidly changing field.

The emergence of more economical technologies for data communications provides, in principle, a method for lowering these geographical and operational barriers; for creating, through computer networking, remote sites at which functionally specialized capabilities are concentrated. The SUMEX-AIM (Stanford University Medical Experimental computer - Artificial Intelligence in Medicine) project is an experiment in reducing this principle to practice, in the specific area of artificial intelligence research applied to health sciences.

The SUMEX-AIM computer facility (3) is a National Shared Computing Resource being developed and operated by Stanford University, in partnership with and with financial support from the Biotechnology Resources Branch of the the Division of Research Resources, National Institutes of Health. It is national in scope in that a major portion of its computing capacity is being made available to authorized research groups throughout the country by means of communications networks.

Aside from demonstrating, on managerial, administrative and technical levels, that such a national computing resource is a viable concept, the primary objective of SUMEX-AIM is the building of a collaborative research community. The aim is to encourage individual participants not only to investigate applications of artificial intelligence in health science, but also to share their programs and discuss their ideas with other researchers. This places a responsibility upon SUMEX-AIM to develop effective means of communication among community members and among the programs they write. It also places responsibility upon those members to design and document programs that readily can be used and understood by others.

Another aspect of the SUMEX facility is providing service to individuals whose interest is in using, rather than developing, the available computer programs. Although this is not a primary consideration, it is an essential part of the growth of these programs. Most of the SUMEX-AIM projects have formed, or are forming, their own user communities which provide valuable "real world" experience. Figure 1 depicts the typical interaction of such a project with its user community, and with other projects. The participation by users in program development is not just restricted to suggestions, but can also include software created by computer-oriented users to satisfy special needs. In some projects, methods are being considered to further promote this kind of participation.

The purposes of this paper are threefold: first, to indicate the range of research projects currently active at SUMEX; second, to describe in detail one of these projects, DENDRAL, which is of particular interest to chemists; and third, to discuss some problems and possible solutions related to networking and community-building.

I. Research Activities at SUMEX-AIM

The community of participants in SUMEX-AIM can be divided geographically into local (i. e., Stanford-based) projects and remote projects, and below is given a brief description of the major representatives of each. Communication with the remote projects is accomplished through one or both of the communications networks shown in Figure 2. In most cases, connection with SUMEX-AIM from these remote sites involves only a local telephone call to the nearest network "node".

The SUMEX-AIM system is itself undergoing constant improvement which deserves to be called research, and thus a third section is

included here to represent system developments.

Remote projects

The Rutgers project. Originating from Rutgers University are several research efforts designed to introduce advanced methods in computer science - particularly in artificial intelligence and interactive data base systems - into specific areas of biomedical research. One such effort involves the development of computer-based consultation systems for diseases of the eye, specifically the establishment of a national network of collaborators for diagnosis and recommendations for treatment of glaucoma by computer. Another project concerns the BELIEVER program, which represents a theory of how people arrive at an interpretation of the social actions of others. SUMEX-AIM provides an excellent medium for collaboration in the development and testing of this theory. The Rutgers project includes, in addition, several fundamental studies in artificial intelligence and system design, which provide much of the support needed for the development of such complex systems.

The DIALOG project. The DIAGNOSTIC LOGIC project, based at the University of Pittsburgh, is a large scale, computerized medical diagnostic system that makes use of the methods and structures of artificial intelligence. Unlike most other computer diagnostic programs, which are oriented to differential diagnosis in a rather limited area, the DIALOG system has been designed to deal with the general problem of diagnosis in internal medicine and currently accesses a medical data base which encompasses approximately fifty percent of the major diseases in internal medicine.

The MISL Project. The Medical Information Systems Laboratory at the University of Illinois (Chicago Circle campus) has been established to explore inferential relationships between analytic data and the natural history of selected eye diseases, both in treated and untreated forms. This project will utilize the SUMEX-AIM resource to build a data base which could then be used as a test bed for the development of clinical decision support algorithms.

Distributed Data-Base System for Chronic Diseases. This project, based at the University of Hawaii, seeks to use the SUMEX-AIM facility to establish a resource sharing project for the development of computer systems for consultation and research, and to make these systems available to clinical facilities from a set of distributed data bases. The radio and satellite links which compose the communication network known as the ALOHANET, in conjunction with the APPANET, will make these programs available to other Hawaiian islands and to remote areas of the Pacific basin. This project could well have a significantly beneficial effect on the quality of health care delivery in these locations.

Modelling of Higher Mental Functions. A project at the University of California at Los Angeles is using the SUMEX-AIM facility to construct, test, and validate an improved version of the computer simulation of paranoid processes which has been

developed. These simulations have clinical implications for the understanding, treatment, and prevention of paranoid disorders. The current interactive version (PARPY) has been running on SUMEX-AIM and has provided a basis for improvement of the future version's language recognition capability.

Local Projects

The Protein Crystallography Project. The Protein Crystallography project involves scientists at two different universities (Stanford and the University of California at San Diego), pooling their respective talents in protein crystallography and computer science, and using the SUMEX-AIM facility as the central repository for programs, data and other information of common interest. The general objective of the project is to apply problem solving techniques, which have emerged from artificial intelligence research, to the well known "phase problem" of x-ray crystallography, in order to determine the three-dimensional structures of proteins. The work is intended to be of practical as well as theoretical value to both computer science (particularly artificial intelligence research) and protein crystallography.

The MYCIN project. MYCIN is an evolving computer program that has been developed to assist physician nonspecialists with the selection of therapy for patients with bacterial infections. The project has involved both physicians, with expertise in the clinical pharmacology of bacterial infections, and computer scientists, with interests in artificial intelligence and medical computing. The MYCIN program attempts to model the decision processes of the medical experts. It consists of three closely integrated components: the Consultation System asks questions, makes conclusions, and gives advice; the Explanation System answers questions from the user to justify the program's advice and explain its methods; and the Rule-Acquisition System permits the user to teach the system new decision rules, or to alter pre-existing rules that are judged to be inadequate or incorrect.

The DENDRAL project. This project, being of particular chemical interest, is described in detail in Section II. Through the SUMEX-AIM facility DENDRAL has gained a growing community of production-level users whose experience with the programs is a valuable guide to further development. Although technically users, some members of this community might better be described as collaborators because they have provided SUMEX-AIM with various special-purpose programs which are of interest to other chemists and which extend the usefulness of the DENDRAL programs.

SUMEX-AIM System Development

Current research activities at SUMEX-AIM are developing along several lines. On a system development level there are ongoing projects designed to make the system more user oriented. Currently,

the system can be expected to provide help to the user who is confused about what is expected in response to a certain prompt. A "?" typed by the user, will, in most cases, provide a list of possible responses from which to choose. Also available in response to typing "HELP" to the monitor is a general help description containing pointers to files likely to be of interest to a new user.

In an effort to facilitate communication between collaborators, a program called CONFER has been developed to provide an orderly method for multiple participant teletype "conference calls". Basically, the program acts as a character processor for all the terminals linked in the conference, accepting input from only one at a time, and passing it out to the remaining terminals. In this way, the conference, in effect, has a "moderator" terminal, thus allowing for a more orderly transfer of ideas and information.

SUMEX-AIM is also aware of the necessity of making its facilities available for trial use by potential users and collaborators. To this end, a GUEST mechanism has been established for persons who wish to have brief, trial access to certain programs they feel may be of value to them, and about which they would like to obtain more knowledge. This provides a convenient mechanism whereby persons, who have been given an appropriate phone number and LOGIN procedure, can dial up SUMEX-AIM and receive actual experience using a program they may only have heard about.

Another area of system development currently being explored at SUMEX-AIM is that of creating a comprehensive "bulletin board" facility where users can file "bulletins", that is, messages of interest to the SUMEX-AIM community. The facility will also alert users to new bulletins which are likely to be of interest to them, as determined by individual user-interest profile.

II. DENDRAL - Chemical Applications of Interactive Computing in a Network Environment

The major research interest of the DENDRAL Project at Stanford University is application of artificial intelligence techniques for chemical inference, focusing in particular on molecular structure elucidation. Portions of our research are in the area of combined gas chromatography/high resolution mass spectrometry and include instrumentation and data acquisition hardware and software development. This area is beyond the scope of this report; we focus instead on the concurrent development of programs to assist chemists in various phases of structure elucidation beyond the point of initial data collection. SUMEX-AIM provides the computer support for development and application of these programs.

Another aspect of our research is our commitment to share developments among a wider community. We feel that several of our programs are advanced enough to be useful to chemists engaged in

related work in mass spectrometry and structure elucidation in general. These programs are written primarily in the programming language INTERLISP, and thus are not easily exportable (exceptions are indicated subsequently). SUMEX-AIM provides a mechanism for allowing others access to the programs without the requirement for any special programming or computer system expertise. The availability of the SUMEX facility over nationwide networks allows remote users to access the programs, in many instances via a local telephone call.

Much of the following discussion is preliminary because our programs have only recently been released for outside use. Some announcement of their availability has been made, and other announcements will occur in the near future, through talks, publications in press, demonstrations and informal discussions. Although most of our experience has been with local users, they have been good models of remote users in that their previous exposure to the actual programs and computer systems is minimal. Their experience has been extremely useful in helping us to smooth out clumsy interactions with programs and to locate and fix program bugs. Such polishing is important for programs which may be utilized by users from widely differing backgrounds with respect to computers, networks and time sharing systems. We are in the processes of building a community of remote users. We actively encourage such use for two reasons: 1) we feel the programs are capable of assisting others in solving certain molecular structure problems, and 2) such experience with outside users will be a tremendous assistance in increasing the power of our programs as the programs are forced to confront new real-world problems.

The remainder of this section outlines the programs which are available via SUMEX, the utilization of these programs in helping to solve structure elucidation problems and the limitations we see to their use. We discuss current applications of the programs to our research and the research of other users to illustrate better the variety of potential applications and to stimulate an interchange of ideas. where appropriate, we point out current difficulties with the use both of our programs and of SUMEX. New applications and wider use will certainly change the nature of these problems; we strive to solve current problems, but new ones will always arise to take their place.

DENDRAL Programs

We have several programs which we employ in dealing with various aspects of problems involving unknown structures. Some of these programs are exportable, while the remainder are available at SUMEX. The availability of each program is discussed below.

Our initial emphasis in studying applications of artificial

intelligence for chemical inference was in the area of mass spectrometry(4-6). This emphasis remains because many of our problems require mass spectrometry as the analytical tool of choice in providing structural information on small quantities of sample. More recently, we have been developing a program (CONGEN, below) directed at more general aspects of structure elucidation. This has extended the scope of problems for which we can provide computer assistance.

We will begin, however, with discussion of the mass spectrometry programs. The examples used in the discussion are characteristic of our current research problems, although we have focused on relatively simple problems to keep the presentation brief. We trace, in what might be chronological terms, the application of the programs to various phases of a structure problem. In this way we hope to illustrate the place of each program in the analysis. We begin by discussing preprocessing of mass spectral data (CLEANUP and MOLION). Subsequent analysis of such data in terms of structure is then covered (PLANNER). The use of CONGEN is discussed for problems which cannot be handled by the previous programs. Finally, we discuss efforts to discover, with the use of the computer, systematics in the behavior of known substances in the mass spectrometer as a means of extending the knowledge of the system for applications in new areas (INTSUM and RULEGEN).

Applications to Molecular Structure Problems

The first three programs, CLEANUP, the library-search program and MOLION are in a sense utility programs, but all three play a critical role in processing mass spectral data. Subsequent applications of programs (e.g., PLANNER) for more detailed spectral analysis in terms of structure depend on the successful treatment of the data by CLEANUP and MOLION, while the library search program filters out common spectra which need not undergo a full analysis. The examples used are drawn from our collaboration with persons in the Genetics Research Center at Stanford Hospital. The experimental data which are collected are the results of combined gas chromatographic/low resolution mass spectral (GC/LRMS) analysis of various fractions (chemically fractionated and derivatized where necessary) of body fluids, e.g. blood, urine. A typical experiment consists of 500-600 individual mass spectra for each fraction, taken sequentially over time as the various components, largely separated from one another, elute from the gas chromatograph and pass into the mass spectrometer. Each mass spectrum consists of the mass analyzed fragment ions of the component(s) in the mass spectrometer at the time the spectrum was taken. Such spectra are related, indirectly, to the molecular structure of the component(s).

CLEANUP(7). The individual mass spectra obtained from fractionated GC/LRMS analysis are quite often poor representations of corresponding spectra taken from pure compounds. They can be

contaminated by the presence of additional peaks and/or distortions to the intensities of existing peaks in the spectrum. Fragment ions from either the liquid phase of the GC column or from components incompletely separated by the gas chromatograph are responsible for the contamination. We have developed a program, referred to here as CLEANUP, which examines all mass spectra in a GC/LRMS run, selects those spectra which contain ions other than background impurities, and remove contributions from background and overlapping components. A spectrum results which compares favorably with the spectrum of a pure component. Biller and Biemann (8) have developed a similar but less powerful program.

For example, the CLEANUP program detected components at points marked with a vertical bar in the plot of total ion current vs. scan number (time), Figure 3. Note that overlapping components were detected under the envelopes of the GC peaks in the region of scans 485-488, 525-529 and 539-552. We focus our attention on the spectrum recorded at scan 492. The raw data, prior to cleanup, are presented in Figure 4 (top). The spectrum resulting from CLEANUP is presented in Figure 4 (bottom). Note that the large ions (e.g., m/e 207, 221 and 315) from background impurities are removed, and that the intensity ratios of peaks at lower masses (e.g., 51 and 77) have been adjusted to reflect their true intensities in the spectrum.

The CLEANUP program is capable of detection of quite low-level components in complex mixtures as indicated by some of the areas of the total ion current plot (Figure 3) where components were detected. It is completely general because nothing in the program code is sensitive to the types of compounds analyzed or the characteristics of possible impurities associated with the compounds or from the GC column. Its major limitation is that mass spectra must be taken repetitively during the course of a GC/MS run. Its performance is enhanced when such spectra are measured closely in time.

The program is offered via SUMEX as an adjunct to use of our other programs; it is not offered as a routine service. Because the program is written in FORTRAN, we routinely use it on our data acquisition computer system so as not to burden SUMEX with tasks better done elsewhere. Similarly, we would assist other frequent users to mount the program on their own systems.

Library Search. With a set of "clean" mass spectra available, the next problem is identification of the various components. Over the course of several years, libraries of mass spectral data have been assembled(9). These libraries can be very useful in weeding out from a group of spectra those which represent known compounds(10). Clearly, one should spend time on solving the structures of unknown compounds, not on rediscovering old ones. The CLEANUP program provides mass spectra which are of sufficient quality to expect that known compounds would be identified easily from such libraries.

Insert A - Dennis' sep. page. This brief example illustrates the obvious value and limitations of library searching. The most interesting compounds for subsequent analysis are those which are unknown. The fractions of urine extracts are replete with unidentified compounds because of the inadequacy of current library compilations. As new compounds are identified they are, of course, added to the library so that future analyses need not reinvestigate the same material.

We currently perform library searching on our data acquisition and reduction computer systems. We can, if necessary, offer limited library search facilities via SUMEX. However, because commercial facilities are available (e.g., over the GE network), routine library search service is not available on SUMEX.

MOLION(11). At this stage we are left with a collection of mass spectra of unknown compounds. The library search results may have provided some clues as to the type of compound present, e.g., compound class. Structure elucidation now begins in earnest. The key elements in problems of structure elucidation are the molecular weight and empirical formula of a compound. Without these essential data, the structural possibilities are usually too immense to proceed further. Mass spectrometry is frequently used to determine molecular weights and formulae, but there is no guarantee that the mass spectrum of a compound displays an ion corresponding to the intact molecule. For example, many of the derivatives of the amino acid fractions of urine display no molecular ions. When we are given only the mass spectrum (and for GC/MS analysis a mass spectrum may be all that is available) we must somehow predict likely molecular ion candidates. The program MOLION performs this task. Given a mass spectrum, it predicts and ranks likely molecular ion candidates independent of the presence or absence of an ion in the spectrum corresponding to the intact molecule. The published manuscript(11) provides many examples of the performance of the program.

The mass spectrum of an example, unknown X, (which we will pursue in more detail below) is given in Figure 5. The results obtained from MOLION are summarized in Table I. The observed ion at m/e 263 is ranked as the most likely candidate.

Table I. Results of Molecular Ion Determination for the Unknown Compound, X, whose Mass Spectrum is Presented in Figure 5.

CANDIDATE	RANKING INDEX
263.0	100
307.0	41
299.0	38

295.0
281.0

34
25

The MOLION program is written to operate on either low or high resolution mass spectra. The program has certain limitations which have been summarized in detail previously(11).

MOLION is available on SUMEX. A FORTRAN version, initially for low resolution mass spectra, is being written so that the program can be run on smaller computers and exported to others. However, it will continue to be available via SUMEX so that others can access it easily. MOLION is contained within PLANNER as one of the available methods for detecting candidate molecular ions.

PLANNER(12). The PLANNER program is designed to analyze the mass spectrum of a compound or of a mixture of related compounds. Because there is no ab initio way of relating a mass spectrum of a complex organic molecule to the structure of that molecule, PLANNER requires fragmentation rules for the class of compounds to which the unknown belongs. This is its major limitation. For our example the class was unknown, forcing us to resort to other means of assistance.

Applications and limitations of PLANNER have been discussed extensively(12,13). The program is very powerful in instances where mass spectrometry rules are strong (i.e., general, with few exceptions). In instances where rules are weak or nonexistent, additional work on known structures and spectra may yield useful rules to make PLANNER applicable (see INTSUM and RULEGEN, below). One useful feature of PLANNER is its ability to analyze the spectra of mixtures in a systematic and thorough way. Thus, it can be applied to spectra obtained as mixtures when GC/MS data are unavailable or impossible to obtain. PLANNER is available in an interactive version over SUMEX, requiring three kinds of information as input: the high or low resolution mass spectrum, the characteristic skeletal structure for molecules in the specific compound class, and the fragmentation rules for the class. Additional knowledge about the unknown can be used by the program to constrain the structural possibilities.

PREDICTOR. The purpose of the DENDRAL predictor is to make a testable prediction for each candidate structure and suggest crucial data points that would allow a chemist to distinguish among the candidate structures. The program is a simulation of the mass spectrometer that predicts both mass spectral peaks and metastable peaks for each candidate. A rudimentary additional program looks for mass spectral ions and metastable ions that are unique for each candidate, thereby providing means of disconfirming or confirming individual candidates.

CONGEN(14,15). Structure problems are usually not solved with mass spectrometry alone. Even when sample size is too limited for obtaining other spectroscopic data, knowledge of chemical isolation and results of derivatization procedures frequently acts as powerful constraints on structural possibilities. Larger amounts of sample permit determination of other spectroscopic data. Taken together, this information allows determination of structural features (substructures) of the molecule and constraints on the plausibility of ways in which the substructures may be assembled. The CONGEN program is capable of providing assistance in solution of such problems.

CONGEN performs the task of construction, or generation, of structural isomers under constraints. The program accepts as input known structural fragments of the molecule ("superatoms") and any remaining atoms (C,N,O,P,...), together with constraints on how they may be assembled. It is based on the exhaustive structure generator(16,17) and extensions(18) which permit a stepwise assembly of structures.

In an interactive session with the program, a user supplies structural information determined by his own analysis of the data (perhaps with the help of the above programs), together with whatever other constraints are available concerning desired and undesired structural features, ring sizes and so forth. The program builds structures in a series of steps, during which a user can interact further with the procedure, for example, to add new constraints. Although very much a developing program, its ability to accept user-inferred constraints from many data sources makes CONGEN a general tool for structure elucidation which we are making available in its current form.

For the unknown X, the observed fragment ions from the molecular ion (M) at m/e 263 (Figure 5) suggest several structural features when coupled with the knowledge of the chemical derivatization procedures used on this fraction of the urine extract. The ion at m/e 194 represents loss of 69 amu, probably CF₃, from fragmentation of a trifluoroacetyl derivative of an amine. This suggests the partial structure 2, Figure 5. The ions at m/e 190 (M-74 amu) and m/e 162 (M-101 amu) suggest the characteristic fragmentation of an n-butyl ester resulting from the second derivatization procedure, formation of the n-butyl esters of free carboxylic acid functions. This suggests the partial structure 1, Figure 5. Taken together, all the above information implies (if no other elements are present) that the empirical formula contains an odd number of nitrogen atoms, at least three oxygen atoms, three fluorine atoms and at least seven carbon atoms. Interestingly, there is only one plausible empirical formula under these constraints, C₁₁H₁₂NO₃F₃.

Structural fragments ("superatoms") 1 and 2 were supplied to CONGEN, together with the remaining four carbon atoms and three degrees of unsaturation (that is, rings plus multiple bonds). With no additional constraints, 155 structures result. The inclusion of other plausible constraints (e.g., no allenes, acetylenes, cyclopropenes, cyclobutenes) reduces the number of structural candidates to just the two isomeric forms of 3, Figure 5.

This problem represents a simple example of a large class of such problems. Although a chemist could probably reach the same conclusions quickly in this case, in the general case, piecing together potential solutions is not a trivial task.

Although still a developing program, CONGEN is, capable of considerable assistance in a wide variety of structure problems. Some areas of current application are summarized in the subsequent section. It is already proving its value in structure elucidation problems by suggesting solutions with a guarantee that no plausible alternatives have been overlooked.

The program has a great deal of flexibility. Many of the types of constraints normally brought to bear on structure elucidation problems can be expressed. However, some types of constraints cannot be easily expressed (e.g., disjunctions of features and stereo-constraints). Recent work by our group and Wipke's(19) will make it possible to add considerations of stereoisomerism relatively easily (a good example of collaboration via SUMEX). We are depending on a broad user community to help us guide further development of CONGEN.

Knowledge Acquisition

INTSUM(20) and RULEGEN. When the mass spectrometry rules for a given class of compounds are not known, the INTSUM and RULEGEN programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the molecules whose spectra display evidence for each particular fragmentation, along with

the total (and average) ion current associated with the fragmentation.

The RULEGEN program attempts to explain the regularities found by INTSUM in terms of the underlying structural features around the bonds in question that seem to "direct" the fragmentations. For example, INTSUM will notice significant fragmentation of the two different bonds alpha to the carbonyl group in aliphatic ketones. It is left to RULEGEN to discover that these are both instances of the same fundamental alpha-cleavage process that can be predicted any time a bond is alpha to a carbonyl group.

These programs are part of the so-called Meta-DENDRAL effort, whose general goal is to understand rule formation activities. Both INTSUM and RULEGEN are available as interactive programs on SUMEX, the former being much more highly developed than the latter. Although these programs can be very useful to chemists interested in finding new mass spectrometry rules, they require having the collection of mass spectra and molecular structure descriptions available in one computer file. Because of this, they have been used mostly by chemists at Stanford.

Applications and Resource Sharing

The DENDRAL programs are being developed to serve a broad community of chemists with structure elucidation problems. Our experience is admittedly limited. In this section we discuss some of the applications, both local and from remote sites, where these programs have proven useful.

CLEANUP. This program has been developed in collaboration with the research staff of the Genetics Research Center at Stanford. Because that staff is working on problems in which few assumptions can be made about the samples, the CLEANUP program has been made very general. For example, the program works with high or low resolution spectra, and makes no assumptions about the actual GC columns used for separating components. This program will be tried next on the GC/HRMS data collected on the MAT-711 in the mass spectrometry laboratory of the Stanford chemistry department.

MOLION. The molecular ion determination program has been used in conjunction with CLEANUP on mass spectrometry problems in the Genetics Research Center. Because of the few assumptions that can be made about the samples in the GRC, the MOLION program has been made very general. We have incorporated this program as part of the PLANNER, as a powerful, class-independent method for inferring the composition of the molecular ion before structure analysis. We anticipate wide use of the program when the FORTRAN version is available for export as well as stand-alone use on SUMEX.

PLANNER. The planning program has been used to infer plausible placement of substituents around a skeletal structure for numerous test problems in which the class of the sample was known and the fragmentation rules for the class were known. Those tests have resulted in a program that we believe is general. We have applied this program to unknown mixtures of estrogenic steroids(13). We are preparing to use PLANNER for screening mass spectra of marine sterols to identify quickly those spectra of known compounds and to suggest structures for spectra of new compounds.

CONGEN. CONGEN is being used locally and from remote sites in a wide variety of applications. We have used it for construction of ring systems under constraints(21) and for generation of structures of chlorocarbons(22). We have investigated several monoterpenoid and sesquiterpenoid structure problems to suggest solutions and to ensure that all alternatives had been considered. We are currently investigating the scope of terpenoid isomerism. Two problems relating to unknown photochemical reaction products have been analyzed and results used to suggest further experiments. In most cases we do not know the precise problems under study by remote users, only that they are using the program.

CONGEN will perhaps be the most widely used (by remote users) program of those mentioned above as accessible through SUMEX. This is primarily a result of the wider scope of problems which might benefit from use of the program. However, the need for remote users to have their mass spectral data available at SUMEX for analysis present a significant energy barrier to use of the programs which require these data.

INTSUM and RULEGEN. INTSUM is essentially a production program now, and is being used as such in a variety of applications involving correlations of molecular structures with their respective spectra. Recent or current applications include analysis of the mass spectra of progesterones and related steroids, androstanes, macrocyclic antibiotics, insect juvenile hormones and phytoecdysones.

These studies serve to develop fragmentation rules which, if of sufficient generality, can in turn be used in PLANNER in the study of unknown compounds.

III. Problems related to networking

During this first year of operation, the SUMEX-AIM facility has encountered a variety of problems arising from its network availability. In most cases, there has been no clear precedent for the handling of these situations, in fact, many problem-areas still reflect the influences of a yet-developing policy. The hope is that this presentation and discussion of problems and their solutions may give foresight to others who contemplate networking or network use.

The problems to be discussed can be loosely associated into three classes; those related to the management of the facility, those pertaining to research activities on the system, and those involving psychological barriers to network use.

Managerial problems

"Gatekeeping." The most general problem faced by the organizers of the SUMEX-AIM facility is the question of "gatekeeping." In order to insure a high quality of pertinent research, some kind of refereeing system is needed to assess the value of proposed new projects. The organizers of the facility would seem to be the best source of such judgements; yet, because we are both organizers and members of the SUMEX community, there is a danger that our decisions would unfairly favor local priorities. In order to establish credibility in SUMEX-AIM as a truly national resource, a management system has been instituted that allocates a defined fraction (initially 50%) of the SUMEX resource to external users, under the jurisdiction of an independent national committee (the AIM advisory group). The remaining 50%, allocated for local use, contains a portion for flexible experiments outside of local projects, but on our own responsibility.

Choice of computer and operating system. A second management level problem is the choice of a computer and operating system which optimize the usefulness of the facility for a majority of users, and which encourage intercommunication between remote collaborators. Because SUMEX-AIM is intended to be used primarily for applications of artificial intelligence, and because interactive LISP (INTERLISP[ref]) is a primary language in this type of work, the choice of TENEX[ref] as an operating system was dictated somewhat by necessity. TENEX incorporates multiple address spaces, thereby allowing multiple "fork" structure and paging, a design which is necessary to create the large-memory virtual machine required by INTERLISP.

The PDP-10 is a popular machine for interactive computing of all sorts in university research environments, and thus an added benefit of this choice was expected - the possibility of easily transferring to SUMEX programs developed at other sites. Many of these programs were written not under TENEX but under the 10/50 monitor supplied by the manufacturer. Because a large and useful program library was already available under the 10/50 monitor, one of the design criteria of TENEX was compatibility with such programs; when a 10/50 program is run under TENEX, a special "compatibility package" of routines is invoked to translate 10/50 monitor calls into equivalent TENEX monitor calls. Although the concept is sound, we have found that in practice very few programs written for the 10/50 monitor are able to run under TENEX without extensive modification. Other problems with TENEX include weaknesses in the support of peripheral devices and the lack of a default line-editor. The latter has caused a proliferation of editing programs, and some confusion has resulted because editor conventions vary from program to program. These difficulties have dampened somewhat

our initial enthusiasm for the TENEX system.

Nonetheless, TENEX provides some features which are crucial to a comfortable network environment. The standard support programs included with this system facilitate both the sending of messages to other users (either at the same site or at other sites on the ARPA network) and the transfer of data and programs from site to site on that network; also, the ability to "link" two or more terminals allows users to communicate easily and immediately. Both the linking and message facility have been found to be invaluable aids in inter-group communications and in such problems as interactive program debugging. When two terminals are linked, their output streams are merged, thus allowing each terminal to display everything typed at the other terminal. Since only the output stream is affected under these circumstances, it is still possible for each terminal to be used to provide input to separate programs, in addition to being used in a conversational mode.

Maintaining a livable system. This third management-level problem is multi-faceted, with major areas of concern being resource allocation, security and file backup. Each of these issues involves keeping the system comfortably useful for the members of the SUMEX community. Although these difficulties are felt at sites which serve only a local community, they are accentuated by network connection.

As noted above, the computational resources of the SUMEX-AIM facility are apportioned by the AIM advisory group and SUMEX management. Some extensions to the basic TENEX system have been made to reflect this apportioning in the actual use of the facility. Basically, it was recognized that users of the facility are members of groups working on specific projects, and it is among these projects that the facility is apportioned. Disk space and cpu cycles are now distributed among groups instead of among individual users. For example, a user may exceed his individual disk allocation somewhat without any ill effect, so long as the total allocation of his group remains within the limits. Similarly, a Reserve Allocation Scheduler has been added to TENEX which tries to match the administrative cycle distribution over a ninety second time frame. Thus a particular group cannot dominate the machine if a lot of its members are logged in at one time.

It is typical for usage of a facility to peak through the middle hours of the day. Indeed, one of the advantages of having users from around the country is the spreading of the load caused by the difference in time zones. Even so, the facility could offer better service if more people would shift their main usage hours toward either end of the day. To encourage "soft-scheduling" within groups on the system, SUMEX-AIM publishes a weekly plot of diurnal loading. This plot shows the total number of jobs on the system as well as the number of LISP jobs, since these jobs seem

to make the biggest demands of system resources. The result has been an increased awareness by users of system loading and a noticeable increase in the number of users at all hours of the night and early morning.

Protection for a computer system covers a range of ideas. It means the ability to maintain secrecy - for example, to guarantee the privacy of patient records. It also guarantees integrity by assuring that programs and data are not modified by an unauthorized party.

Questions of protection generally become more interesting and complex as more sharing is involved. Consider the example of a proprietary program which generates layouts given a user's circuit data. The program owner demands assurance that he will be paid whenever his program is used and that copies of the program cannot be made. The user wants guarantees that his data sets cannot be destroyed or copied for a competitor. yet the user must have access to the program and the program must have access to the data. Unable to support such complicated examples of protection, SUMEX-AIM assumes that sharing takes place between friendly users. This is not to imply that issues of protection and sharing have not appeared. For example, in an effort to improve the human engineering of programs for public use, the capability of recording a session has been built into several of the programs. Studied by the program designers to pinpoint confusing aspects of programs, these recordings serve to improve program design. Since the issue of violation of privacy has been raised, some of these programs now request permission to record a session before doing so. At this time, any guarantee of privacy must be provided by the program designer because TENEX itself does not have the ability to render the protection .

The general design for systems offering "state of the art" protection involves a tolerance for failure; that is, if a potential offender succeeds in breaking through some of the defenses, he still does not place the entire computer system at his mercy. Encrypting of data files provides an additional line of defense. This method is used by at least two calendar or appointment programs on the computer. At this time, however, there are no general encrypting facilities available and users must do this for themselves as needed.

Tenex provides the usual keyword protection at login time and a measure of file protection. Owners of a file may assign a protection number which specifies some combination of READ, WRITE, EXECUTE, or APPEND access to a file for owners, members of a group, or other users. This level of protection is basically enough to prevent accidents and most mischief. System programmer's around the country are aware of a number of TENEX bugs which permit this access to be violated. One user of our system found a way to place himself in a

mode where he could modify any file on the system. To date, we have no examples of such activity~ actually having a deleterious effect on SUMEX-AIM.

To make the use of SUMEX-AIM programs easily available on a trial basis for prospective users, a "guest" account system has been established. Since this makes logging into SUMEX-AIM so easy, it has invited some misuse by people using those accounts to play the computer games. A proposed extension to the system now being implemented is a special "guest EXEC" which would extend the protection of the TENEX monitor by allowing guest accounts access to only a more restricted set of programs.

In order to assure the user maximum protection against loss of valuable work, SUMEX operates a multi-level file backup system. In addition to routine file backup system there are facilities to enable the user to selectively archive his or her disk files. By issuing a simple command to the TENEX executive the user can transmit a message to the operator to copy specified files to magnetic tape. Each such file is copied to two magnetic tapes within 24 hours of issuing the archive command. File retrieval is affected by a similar process. The user also has the alternative option of being able to lodge files in a special backup directory. Files are held in this directory until the next exclusive file dump (see below) at which time they are deleted. In this way the user can remove files from his directory at his own choosing knowing they will be archived by the exclusive dump.

On a system level, an effort is made to maintain file backups such that the maximum possible loss, in the event of a crash fatal to the file system, would amount to no more than one day's work. Once each day all files that have been read or written within the last 48 hours are dumped onto magnetic tape. Files that exist for 48 hours are thus held on two separate tapes. The rotation period for files dumped in this way is 60 days. Once each week a full file dump is made to separate disk storage. Each such dump is kept for two weeks at which time it is replaced by a new file dump. Each month there is a full system dump from disk to magnetic tape. Files can be recovered from the system backup by sending a message to the operator specifying the file name(s) and when the file was last read or written (if such information is available).

Excessive demand for production programs. One of the concepts behind the creation of a shared resource is elimination of the problems which arise when large, complex computer programs are exported. Since, in theory, exportability is no longer a problem, there is greater latitude in choice of a language in which program development can take place. In the case of some of the DENDRAL programs, it was thought that program development should take place in INTERLISP, a language that lends itself well to the artificial intelligence nature of these programs, but does not lead to particularly efficient run-time code.

In order to ascertain the usefulness of these programs and to determine what areas remained in need of work, chemist collaborators were sought. As these users increased in number and began to use the programs more frequently, it became obvious that the inherent slowness of the predominately LISP code was affecting the whole system as well as handicapping the efficient use of the DENDRAL programs. Additionally, some of the chemist-users who were finding the programs most useful and who were most enthusiastic about their potential use, were persons who were working in industry. Although, in one sense, this interest from industry could be interpreted as an indication of the "real-world" usefulness of the programs, it came as rather a surprise to both SUMEX and DENDRAL personnel.

The fact that SUMEX-AIM is funded by NIH as a national resource prohibits the facility from providing a service, at taxpayer's expense, to a private industry. Although there is precedent for a site funded via government grant to charge a fee for service, such an arrangement leads to highly complicated bookkeeping, and is contrary to the essential purpose of SUMEX-AIM; to be a research-oriented rather than service-oriented facility. This leaves the industrial users in the position of being more than willing to pay for the use of the programs, but of having no mechanism whereby they can be charged. Furthermore, the fact that the programs are coded in LISP for a highly specialized environment, almost guarantees the impossibility of export, except to an almost identical computer system.

An intermediate solution that will help to solve the problem of industrial users on SUMEX and will help to alleviate the system loading resulting from heavy usage of LISP coded production programs, is to mount CONGEN on a closely related computer which is operated on a fee for service basis. However, in order to make this transfer economically feasible, it has become evident that it will be necessary to recode the LISP sections of the program into a more efficient and easily exportable language.

Research-oriented problems

Community mindedness. Those involved in computer science research at SUMEX face a general problem which is absent or greatly lessened at non-network sites; the problem of community mindedness. The network provides a large and varied set of other researchers and users who have an interest in their work. Although the network-TENEX combination provides new forms of communication with these remote parties, the traditional means of fully describing the use and structure of a complex program, a detailed person-to-person discussion, is not convenient. Comprehensive documentation gains importance in such a situation, and within the DENDRAL project a great deal of time has been needed in the development of program descriptions which are adequate for a diverse audience. Also, in both DENDRAL and MYCIN, effort has been and is being directed toward "human engineering" in program design; to provide the user with commands which assist him in using

the programs, in understanding the logic by which the programs reach certain decisions and in communicating questions or comments on the programs' operation to those responsible for development. Such "housekeeping" tasks can often be neglected, yet are quite important in smoothing interaction with the community.

Choice of programming language. High level programming languages which are designed for ease of program development are frequently poor as production-level languages. This is because developmental languages free the researcher from a raft of programming details, thus allowing him to concentrate upon the central logical issues of the problem, but the automatic handling of these details is seldom optimal. Also, because such languages tend to be specialized for certain computers and operating systems, the exportation of programs can be a serious problem. One solution to these problems is the recoding of research-level programs into more efficient language when fast and exportable versions are needed.

Networking greatly eases the problem of exportability, but can also aggravate the the problem of efficiency. As mentioned in the previous section, the DENDRAL programs, which are undergoing constant development, found a substantial number of production-level users. Because of the inefficiencies of INTERLISP (a 50- to 100-fold improvement in running time is not uncommon when an INTERLISP program is translated into FORTRAN), this use adversely impacted the entire system. Because the DENDRAL programs are quite large and complex, their translation into other languages is impractically tedious. A partial solution to this problem is provided by the TENEX operating system, which allows some interface between programs written in different languages. With such intercommunication, time-consuming segments of an INTERLISP program which are not undergoing active development can be reprogrammed in another more efficient language. The developmental parts of the program are left in INTERLISP, where modifications can easily be made and tested. The CONGEN program uses three languages; INTERLISP, FORTRAN and SAIL[ref]. The SAIL segment was added when a new feature, whose implementation was fairly straightforward, was included in CONGEN. Since then, the SAIL portion gradually has been taking over some of the more time-consuming tasks. This method allows a balance in the tradeoff between ease of program development and efficiency of the final program.

Accumulation of expert knowledge in knowledge-based programs. Just as statistics-based programs need to worry about accumulation of large data bases, knowledge based programs need to worry about the accumulation of large amounts of expertise. The performance of these programs is tied directly to the amount of knowledge they have about the task domain -- in a phrase, knowledge is power. Therefore, one of the goals of artificial intelligence research is to build systems that not only perform as well as an expert but that also can accumulate knowledge from several experts.

Simple accretion of knowledge is possible only when the "facts", or inference rules, that are being added to the program are entirely separate from one another. It is unreasonable to expect a body of

knowledge to be so well organized that the facts or rules do not overlap. (If it were so well organized, it is unlikely that an artificial intelligence program would be the best encoding of the problem solver.) One way of dealing with the overlap is to examine the new rules on an individual basis, as they are added to the system in order to remove the overlap. This was the strategy for developing the early DENDRAL programs. However, it is very inefficient and becomes increasingly more difficult as the body of knowledge grows.

The problem of removing conflicts, or potential conflicts, from overlapping rules becomes more acute when more than one expert adds new rules to the knowledge base. Of course, the advantages of allowing several experts to "teach" the system are enormous -- not only is the program's breadth of knowledge potentially greater than that of a single expert, but the rules are more apt to be refined when looked at by several experts. On the other hand, one can expect not only a greater volume of new rules but a higher percentage of

conflicts when several experts are adding rules.

Having a computer program that can accumulate knowledge presupposes having an organization of the program and its knowledge base that allows accumulation. If the knowledge is built into the program as sequences of low-level program statements -- as often happens -- then changing the program becomes impossible. Thus current artificial intelligence research stresses the importance of separating problem-solving knowledge from the control structure of the program that uses that knowledge.

Another problem, at a political rather than a programming level, becomes apparent with one accumulation process: how does the program distinguish an expert from a novice? In the MYCIN program we have circumvented the problem by having the program ask the current user for a keyword that would identify him as an expert. It is then a bureaucratic decision as to which users are given that keyword. There is nothing subtle in this solution, and one can imagine far better schemes for accomplishing the same thing. The point here is that not every user should have the privilege of changing rules that experts have added to the system, and that some safeguards must be implemented.

"Human nature" barriers to SUMEX use

Countering disbelief. There is sometimes a tendency among those unfamiliar with the capabilities and limitations of computers and computer programs to express disbelief. This is not disbelief in the sense of worrying that the programs have errors and produce erroneous results. Indeed, the fact that a problem is being done by a computer seems to generate some faith that it might be right, or at least significantly reduces questions about correctness. The disbelief is that programs, which are designed to model, or to emulate, human problem solving will not be capable of useful performance. This, of course, is the classic argument against artificial intelligence -- we think in mysterious ways and have such a complex brain that a computer program must be inferior. In some cases, authors of artificial intelligence programs have brought such criticism upon themselves by not stressing limitations, or by making extravagant claims.

In the DENDRAL project, we have tried to counter this type of disbelief in a number of ways. We have tried to stress that our programs are designed to assist, not replace chemists. We have always discussed limitations to give a reasonable perspective on capabilities vs. limitations of a program. Most importantly however, we have focused on those aspects of problems which are amenable to systematic analysis, i.e., those problems which can be done manually, but only with difficulty and with the consumption of a great deal of time which a chemist could better spend on more productive pursuits. Examples of this would include the application of PLANNER to mixtures where all fragmentations may have to be considered as possible fragments of every molecular ion, the systematic analysis by INTSUM

of possible fragmentation processes, the consideration by MOLION of all plausible possibilities, and the structure generation capabilities of CONGEN.

We have also tried to reduce chemists' disbelief by blurring the "outsider-insider" distinction, in particular by having trained chemists work on the programs and make them useful to themselves first. Further, when "outside" chemists are first introduced to the programs, the introduction is done by another chemist who has already thought through and can readily explain many of the chemistry-related problems.

The ultimate way to counter disbelief, however, is to illustrate high levels of performance. If a potential user is aware of the goals (intent) of a program and its limitations, a few examples of results which would be extremely difficult to obtain without the program are very convincing.

The "security" of a local facility. Networking is still a relatively new concept to many people, and there is a resistance to departing from the "traditional" modes of computing. There is a sense of security in having a local computing facility with knowledgeable consultants within walking distance, and in having "hard" forms of input (eg, boxes of computer cards) and output (eg, voluminous listings). These props are difficult to simulate over a network connection - in most cases a user's interaction with the remote site takes place exclusively through a computer terminal - yet the quality of service can match or exceed that of a local facility; programs and large data sets can be entered and stored on secondary storage as can large output files; all types of program and data editing can be done with interactive editing programs; programs can be written in an interactive mode so that small amounts of control information can be input and key results output in "real time" over the terminal; And as noted in a previous section, consultation can be significantly more productive providing that the remote operating system supports the appropriate types of communication possibilities.

There can, of course, be no denying that there are problems in learning to use a distant computer system, be it for program development or for the use of certain programs. Whether or not overcoming these problems to gain access to the special resources which are available, is worth the effort, is a question answerable only by the individuals involved. Fortunately, there will always be those persons who have a pressing problem in need of solution and who are willing to try a new approach; regardless of whether or not they have had prior network experience.

The SUMEX-AIM Facility

The SUMEX-AIM computer facility consists of a Digital Equipment Corporation model KI-10 central processor operating under the TENEX time sharing monitor. It is located at Stanford University Medical Center, Stanford, California.

The system has 256K words (36 bit) of high speed memory; 1.6

million words of swapping storage; 70 million words of disk storage; two 9-track, 800 bpi industry compatible tape units; one dual DEC-tape unit; a line printer; and communications network interfaces providing user terminal access via both TYMNET and ARPANET.

Software support has evolved, and will continue to evolve, based on user research goals and requirements. Major user languages currently include INTERLISP, SAIL, FORTRAN-10, BLISS-10, BASIC and MACRO-10. Major software packages available include OMNIGRAPH, for graphics support of multiple terminal types, and MLAB, for mathematical modelling.

The SUMEX-AIM computer generally is left with no operator in attendance; thereby helping to eliminate some overhead, but also creating some problems. Users who wish to run jobs requiring tapes must make arrangements to mount their own tapes. Likewise, obtaining listings from the line printer can be somewhat difficult since there is no regular schedule for distribution of this output. The solution to these two problems has been to make keys to the machine room available at strategic locations, convenient to all groups of local users. This experiment in basic "resource sharing" has not resulted in any of the major problems one might expect from having a fairly large group of people with hands-on access to a computer.

REFERENCES

1-3 TO BE ADDED

- (4) J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield, and C. Djerassi, J. Amer. Chem. Soc., 91, 2973 (1969).
- (5) A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum, and J. Lederberg, J. Amer. Chem. Soc., 91, 2977(1969).
- (6) B.G. Buchanan, A.M. Duffield and A.V. Robertson, "Mass Spectrometry: Techniques and Applications", G.W.A. Milne, Ed., John Wiley and Sons, 1971, p.121.
- (7) R.G. Dromey, unpublished results, preprint available on request, Dept. of Computer Science, Serra House, Stanford University, Stanford, Calif. 94305.
- (8) J.E. Biller and K. Biemann, Anal. Lett., 515 (1974).
- (9) Several libraries of mass spectral data are available in various forms. The Aldermaston Data Centre (see the "Mass Spectrometry Bulletin") can provide information on the availability of such libraries.
- (10) H.S. Hertz, R.A. Hites, and K. Biemann, Anal., Chem. 43, 681, (1971).

- (11) R.G. Dromey, B.G. Buchanan, D.H. Smith, J. Lederberg, and C. Djerassi, *J. Org. Chem.*, 40, 770 (1975).
- (12) D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg and C. Djerassi, *J. Amer. Chem. Soc.* 94, 5962 (1972).
- (13) D.H. Smith, B.G. Buchanan, R.S. Engelmores, H. Adlercreutz, and C. Djerassi, *J. Amer. Chem. Soc.*, 95, 6078 (1973).
- (14) D.H. Smith and R.E. Carhart, Abstracts, 169th Meeting of the American Chemical Society, Philadelphia, April 6-11, 1975
- (15) R.E. Carhart, D.H. Smith, H. Brown and C. Djerassi, *J. Amer. Chem. Soc.*, submitted for publication.
- (16) L.M. Masinter, N.S. Sridharan, J. Lederberg and D.H. Smith, *J. Amer. Chem. Soc.*, 96, 7702 (1974)
- (17) L.M. Masinter, N.S. Sridharan, R.E. Carhart and D.H. Smith, *J. Amer. Chem. Soc.*, 96, 7714 (1974).
- (18) H. Brown, *SIAM Journal of Computing*, submitted for publication.
- (19) W.T. Wipke and T.M. Dyott, *J. Amer. Chem. Soc.*, 96, 4825 (1974).
- (20) D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, *Tetrahedron*, 29, 3117(1973).
- (21) R.E. Carhart, D.H. Smith, and H. Brown, *J. Chem. Inf. Comp. Sci.*, in press (May, 1975).
- (22) D.H. Smith, *Anal. Chem.*, in press (May 1975).

Figure Captions

Figure 1. Interactions in the SUMEX-AIM community

Figure 2. Access to SUMEX-AIM

Figure 3. Total ion current vs. spectrum number in a GC/LRMS run

Figure 4. The spectrum corresponding to scan 492 in Figure 3. (top) Raw data. (bottom) Output from CLEANUP

Figure 5. Low resolution mass spectrum of unknown X. The indicated superatoms were deduced from the spectrum and a knowledge of the chemical history of the sample. With these and other constraints, CONGEN obtained the indicated results.

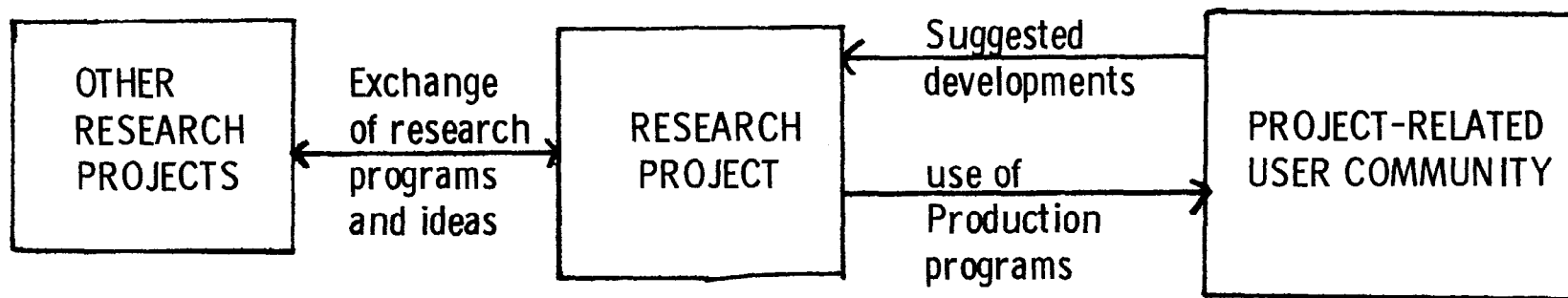


Fig. 1

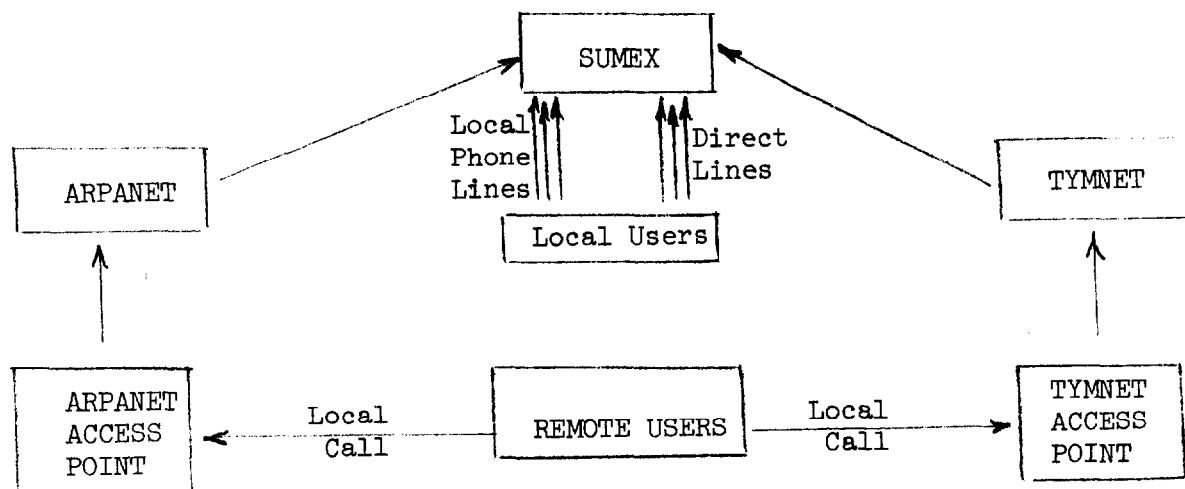


Fig. 2.

Figure 3. GC Trace (to be supplied)

Figure 4. Two Mass Spectra (to be supplied)

Figure 5. Mass Spec of Unknown (to be supplied)

Figure 3

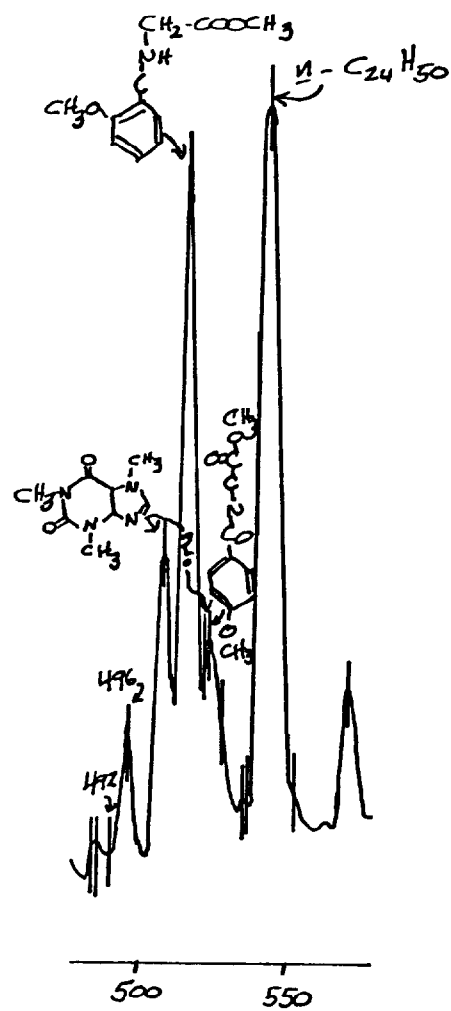
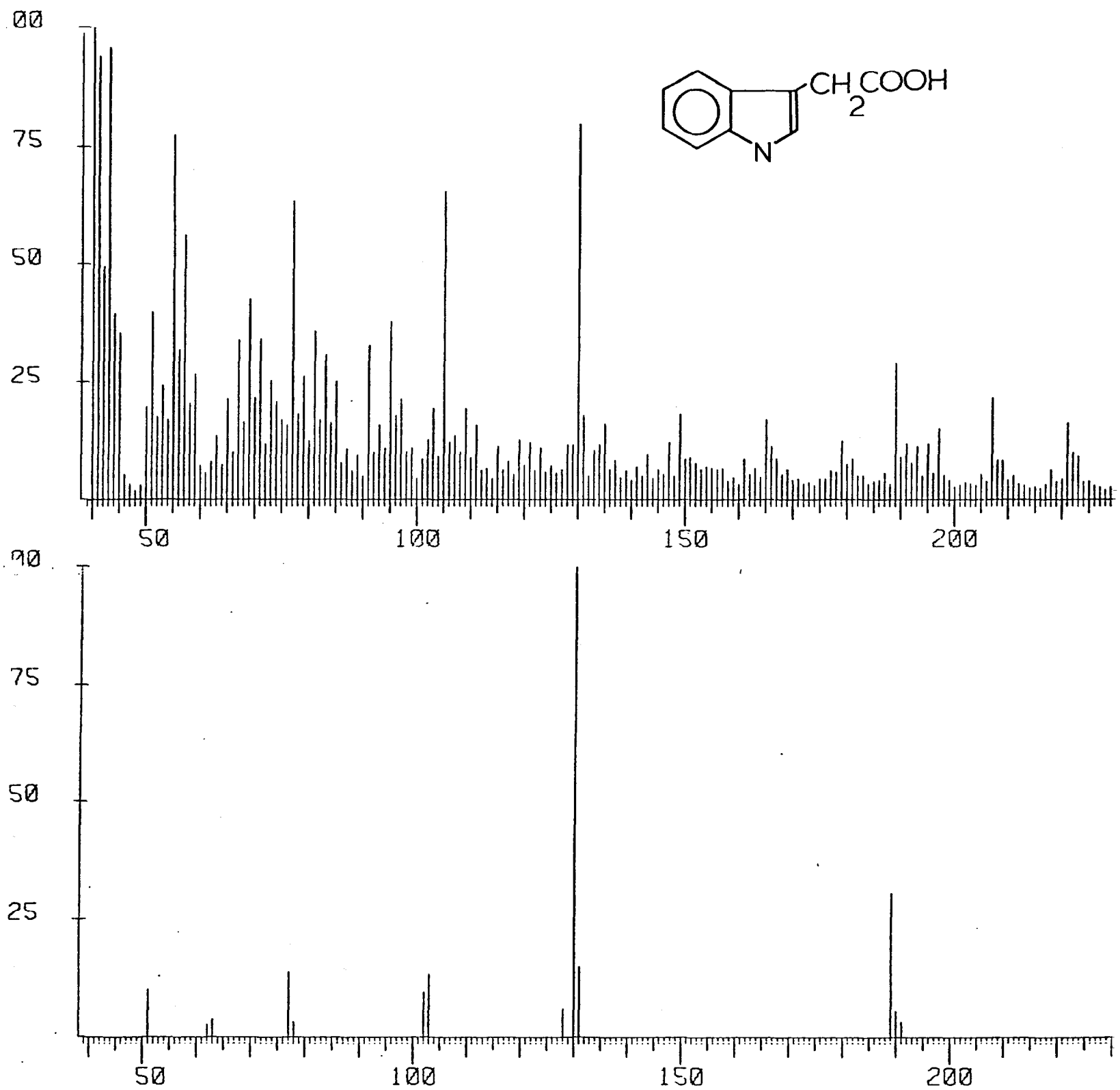


Figure 4

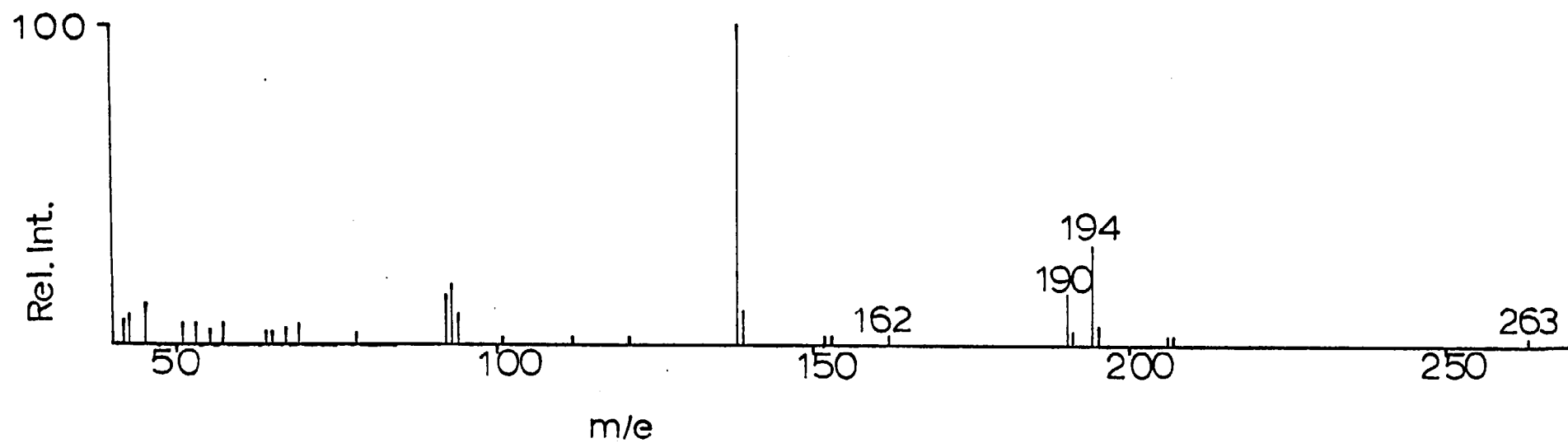


CLEAN SPECTRUM

EXP. W247 20-DEC-74

SPECTRUM NO. 492

Figure 5

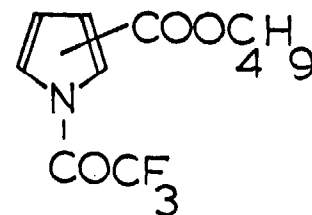


SUPERATOMS

$n\text{-C}_4\text{H}_9\text{OOC-}$ 1

$\text{CF}_3\text{CO-}$ 2

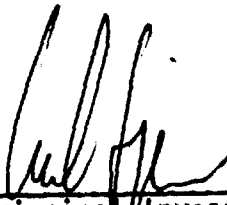
CONGEN RESULT



3

The undersigned agrees to accept responsibility for the scientific and technical conduct of the project and for provision of required progress reports if a grant is awarded as the result of this application.

5/7/75
Date



Principal Investigator or
Project Director

11.B. SUMMARY OF RESOURCE USAGE

The outside uses of our resource-related research are listed in Part III of the Description of Progress (section 11-A).

11.C. RESOURCE RELATED RESEARCH EQUIPMENT LIST

EQUIPMENT SUMMARY

1) MM11-U Memory Module PDP 11/45 CM Central Processor (Ser. 5200) FP11-B Floating Point Processor TM11-EA Nine Channel Magnetic Tape Drive and Controller	\$45,372
2) Systems Industries PDP Model 20 or 45 compatible disk system Model 3040 Controller Daisy Chain Option Certified Disk Pack	11,622
3) GT 40AA Display System 115V	14,359
4) M792 32 word Diode Memory for PDP 11 (4)	1,359
5) Disk Pack	268
6) E-30-2004 RB Bud Cabinet	<u>322</u>
	\$73,302

D. SUMMARY OF PUBLICATIONS

- (1) J. Lederberg, "DENDRAL-64 - A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs", (technical reports to NASA, also available from the author and summarized in (12)). (1a) Part I. Notational algorithm for tree structures (1964) CR.57029 (1b) Part II. Topology of cyclic graphs (1965) CR.68898 (1c) Part III. Complete chemical graphs; embedding rings in trees (1969)
- (2) J. Lederberg, "Computation of Molecular Formulas for Mass Spectrometry", Holden-Day, Inc. (1964).
- (3) J. Lederberg, "Topological Mapping of Organic Molecules", Proc. Nat. Acad. Sci., 53:1, January 1965, pp. 134-139.
- (4) J. Lederberg, "Systematics of organic molecules, graph topology and Hamilton circuits. A general outline of the DENDRAL system." NASA CR-48899 (1965)
- (5) J. Lederberg, "Hamilton Circuits of Convex Trivalent Polyhedra (up to 18 vertices), Am. Math. Monthly, May 1967.
- (6) G. L. Sutherland, "DENDRAL - A Computer Program for Generating and Filtering Chemical Structures", Stanford Artificial Intelligence Project Memo No. 49, February 1967.
- (7) J. Lederberg and E. A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry", in B. Kleinmuntz (ed) Formal Representations for Human Judgment, (Wiley, 1968) (also Stanford Artificial Intelligence Project Memo No. 54, August 1967).
- (8) J. Lederberg, "Online computation of molecular formulas from mass number." NASA CR-94977 (1968)
- (9) E. A. Feigenbaum and B. G. Buchanan, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry", in Proceedings, Hawaii International Conference on System Sciences, B. K. Kinariwala and F. F. Kuo (eds), University of Hawaii Press, 1968.
- (10) B. G. Buchanan, G. L. Sutherland, and E. A. Feigenbaum, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry". In Machine Intelligence 4 (B. Meltzer and D. Michie, eds) Edinburgh University Press (1969), (also Stanford Artificial Intelligence Project Memo No. 62, July 1968).
- (11) E. A. Feigenbaum, "Artificial Intelligence: Themes in the Second Decade". In Final Supplement to Proceedings of the IFIP68 International Congress, Edinburgh, August 1968 (also Stanford Artificial Intelligence Project Memo No. 67, August 1968).
- (12) J. Lederberg, "Topology of Molecules", in The Mathematical Sciences - A Collection of Essays, (ed.) Committee on Support of Research in the Mathematical Sciences (COSRIMS), National Academy of Sciences - National Research Council, M.I.T. Press, (1969), pp. 37-51.

- (13) G. Sutherland, "Heuristic DENDRAL: A Family of LISP Programs", to appear in D. Bobrow (ed), LISP Applications (also Stanford Artificial Intelligence Project Memo No. 80, March 1969).
- (14) J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing C, H, O and N". Journal of the American Chemical Society, 91:11 (May 21, 1969).
- (15) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference II. Interpretation of Low Resolution Mass Spectra of Ketones". Journal of the American Chemical Society, 91:11 (May 21, 1969).
- (16) B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry", in Machine Intelligence 5, (B. Meltzer and D. Michie, eds) Edinburgh University Press (1970), (also Stanford Artificial Intelligence Project Memo No. 99, September 1969).
- (17) J. Lederberg, G. L. Sutherland, B. G. Buchanan, and E. A. Feigenbaum, "A Heuristic Program for Solving a Scientific Inference Problem: Summary of Motivation and Implementation", Stanford Artificial Intelligence Project Memo No. 104, November 1969.
- (18) C. W. Churchman and B. G. Buchanan, "On the Design of Inductive Systems: Some Philosophical Problems". British Journal for the Philosophy of Science, 20 (1969), pp. 311-323.
- (19) G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference III. Aliphatic Ethers Diagnosed by Their Low Resolution Mass Spectra and NMR Data". Journal of the American Chemical Society, 91:26 (December 17, 1969).
- (20) A. Buchs, A. M. Duffield, G. Schroll, C. Djerassi, A. B. Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Applications of Artificial Intelligence For Chemical Inference. IV. Saturated Amines Diagnosed by Their Low Resolution Mass Spectra and Nuclear Magnetic Resonance Spectra", Journal of the American Chemical Society, 92, 6831 (1970).
- (21) Y.M. Sheikh, A. Buchs, A.B. Delfino, G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference V. An Approach to the Computer Generation of Cyclic Structures. Differentiation Between All the Possible Isomeric Ketones of Composition C₆H₁₀O", Organic Mass Spectrometry, 4, 493 (1970).
- (22) A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, "Applications of Artificial

Intelligence for Chemical Inference VI. Approach to a General Method of Interpreting Low Resolution Mass Spectra with a Computer", Chem. Acta Helvetica, 53, 1394 (1970).

- (23) E.A. Feigenbaum, B.G. Buchanan, and J. Lederberg, "On Generality and Problem Solving: A Case Study Using the DENDRAL Program". In Machine Intelligence 6 (B. Meltzer and D. Michie, eds.) Edinburgh University Press (1971). (Also Stanford Artificial Intelligence Project Memo No. 131.)
- (24) A. Buchs, A.B. Delfino, C. Djerassi, A.M. Duffield, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, G. Schroll, and G.L. Sutherland, "The Application of Artificial Intelligence in the Interpretation of Low-Resolution Mass Spectra", Advances in Mass Spectrometry, 5 (1971), 314.
- (25) B.G. Buchanan and J. Lederberg, "The Heuristic DENDRAL Program for Explaining Empirical Data". In proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia (1971). (Also Stanford Artificial Intelligence Project Memo No. 141.)
- (26) B.G. Buchanan, E.A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145.)
- (27) Buchanan, B. G., Duffield, A.M., Robertson, A.V., "An Application of Artificial Intelligence to the Interpretation of Mass Spectra", Mass Spectrometry Techniques and Appliances, Edited by George W. A. Milne, John Wiley & Sons, Inc., 1971, p. 121-77.
- (28) D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference VIII. An approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", Journal of the American Chemical Society, 94, 5962-5971 (1972).
- (29) B.G. Buchanan, E.A. Feigenbaum, and N.S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In Machine Intelligence 7, Edinburgh University Press (1972).
- (30) Lederberg, J., "Rapid Calculation of Molecular Formulas from Mass Values". Jnl. of Chemical Education, 49, 613 (1972).
- (31) Brown, H., Masinter L., Hjelmeland, L., "Constructive Graph Labeling Using Double Cosets". Discrete Mathematics, 7 (1974), 1-30. (Also Computer Science Memo 318, 1972).
- (32) B. G. Buchanan, Review of Hubert Dreyfus' "What Computers Can't Do: A Critique of Artificial Reason", Computing Reviews (January, 1973). (Also Stanford Artificial Intelligence Project Memo No. 181)
- (33) D. H. Smith, B. G. Buchanan, R. S. Engelmores, H. Aldercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical

Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". Journal of the American Chemical Society 95, 6078 (1973).

- (34) D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. Intsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". Tetrahedron, 29, 3117 (1973).
- (35) B. G. Buchanan and N. S. Sridharan, "Rule Formation on Non-Homogeneous Classes of Objects". In proceedings of the Third International Joint Conference on Artificial Intelligence (Stanford, California, August, 1973). (Also Stanford Artificial Intelligence Project Memo No. 215.)
- (36) D. Michie and B.G. Buchanan, "Current Status of the Heuristic DENDRAL Program for Applying Artificial Intelligence to the Interpretation of Mass Spectra". August, 1973. To appear in Computers for Spectroscopy (ed. R.A.G. Carrington) London: Adam Hilger. Also: University of Edinburgh, School of Artificial Intelligence, Experimental Programming Report No. 32 (1973).
- (37) H. Brown and L. Masinter, "An Algorithm for the Construction of the Graphs of Organic Molecules", Discrete Mathematics, 8(1974), 227. (Also Stanford Computer Science Dept. Memo STAN-CS-73-361, May, 1973)
- (38) D.H. Smith, L.M. Masinter and N.S. Sridharan, "Heuristic DENDRAL: Analysis of Molecular Structure," Proceedings of the NATO/CNNA Advanced Study Institute on Computer Representation and Manipulation of Chemical Information (W. T. Wipke, S. Heller, R. Feldmann and E. Hyde, eds.) John Wiley and Sons, Inc., 1974.
- (39) R. Carhart and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XI: The Analysis of C13 NMR Data for Structure Elucidation of Acyclic Amines", J. Chem. Soc. (Perkin II), 1753 (1973).
- (40) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Application of Artificial Intelligence for Chemical Inference XII: Exhaustive Generation of Cyclic and Acyclic Isomers". Journal of the American Chemical Society, 96 (1974), 7702. (Also Stanford Artificial Intelligence Project Memo No. 216.)
- (41) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XIII. Labeling of Objects having Symmetry". Journal of the American Chemical Society, 96 (1974), 7714.
- (42) N.S. Sridharan, Computer Generation of Vertex Graphs, Stanford CS Memo STAN-CS-73-381, July, 1973.
- (43) N.S. Sridharan, et.al., A Heuristic Program to Discover Syntheses for Complex Organic Molecules, Stanford CS Memo STAN-CS-73-376, June, 1973. (Also Stanford Artificial Intelligence Project Memo No. 205.)

- (44) N.S. Sridharan, Search Strategies for the Task of Organic Chemical Synthesis, Stanford CS Memo STAN-CS-73-391, October, 1973. (Also Stanford Artificial Intelligence Project Memo No. 217.)
- (45) R. G. Dromey, B. G. Buchanan, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra". Journal of Organic Chemistry, 40 (1975), 770.
- (46) D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XV. Constructive Graph Labelling Applied to Chemical Problems. Chlorinated Hydrocarbons". Analytical Chemistry, in press (to appear May or June, 1975).
- (47) R. E. Carhart, D. H. Smith, H. Brown and N. S. Sridharan, "Applications of Artificial Intelligence for Chemical Inference. XVI. Computer Generation of Vertex Graphs and Ring Systems". Journal of Chemical Information and Computer Science (formerly Journal of Chemical Documentation), in press (to appear in May, 1975).
- (48) R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure". Journal of the American Chemical Society, submitted for publication.
- (49) B. G. Buchanan, "Scientific Theory Formation by Computer." To appear in Proceedings of NATO Advanced Study Institute on Computer Oriented Learning Processes, 1974, Bonas, France.
- (50) E. A. Feigenbaum, "Computer Applications: Introductory Remarks," in Proceedings of Federation of American Societies for Experimental Biology, Vo. 33, No. 12 (Dec., 1974) 2331-2332.
- (51) S. Hammerum and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems - CCXLIV; The Influence of Substituents and Stereochemistry on the Mass Spectral Fragmentation of Progesterone." Tetrahedron (accepted for publication), 1975.
- (52) S. Hammerum and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems CCXLV. The Electron Impact Induced Fragmentation Reactions of 17-Oxygenated Progesterones." Steroids (submitted for publication).
- (53) H. Brown, "Molecular Structure Elucidation III." Submitted for publication to SIAM Journal on Computing.

- (54) W.E. Pereira, R.E. Summons, T.C. Rindfleisch and A.M. Duffield, "The Determination of Ethanol in Blood and Urine by Mass Fragmentography." Clin. Chim. Acta, 51, 109 (1974).
- (55) W.E. Pereira, R.E. Summons, T.C. Rindfleisch, A.M. Duffield, B. Zeitman and J.G. Lawless, "Stable Isotope Mass Fragmentography: Quantitation and Hydrogen-Deuterium Exchange Studies of Eight Murchison Meteorite Amino Acids." Geochem. et Cosmochim. Acta, 39, 163 (1975).
- (56) S.A. Fernback, R.E. Summons, W.E. Pereira and A.M. Duffield, "Metabolic Studies of Transient Tyrosinemia in Premature Infants." Pediatric Research, 9, 172 (1975).
- (57) J.G. Lawless, B. Zeitman, W.E. Pereira, R.E. Summons and A.M. Duffield, "Dicarboxylic Acids in the Murchison Meteorite." Nature, 251, 40 (1974).

III. RESOURCE FINANCES

A. SUMMARY OF EXPENDITURES

B. DETAILS

C. SUMMARY OF RESOURCE FUNDING

D. BUDGET EXPLANATION/JUSTIFICATION

SECTION II

SECTION II—BUDGET (USUALLY 12 MONTHS)	FROM	THROUGH	GRANT NUMBER
	8/1/75	7/31/76	RR612-05A1

A. ITEMIZE DIRECT COSTS REQUESTED FOR NEXT BUDGET PERIOD

PERSONNEL		TIME OR EFFORT %/HRS. (c)	SALARY REQUESTED (d)	FRINGE BENEFITS (See Instructions) (e)	TOTAL (f)
NAME (Last, First, Initial) (a)	TITLE OF POSITION (b)				
	PRINCIPAL INVESTIGATOR				
See separate listing					
		Subtotals	\$	\$	

(Indicate cost of each item listed below)

TOTAL (Columns (d) and (e))

\$ 207,822

CONSULTANT COSTS (See Instructions)

\$ 0

EQUIPMENT PDP-11 Maintenance Contract - \$8,800
MAT-711 Maintenance - \$7,000

\$ 15,750

SUPPLIES Office supplies (500), electronics supplies (1000), GC supplies (1100), liquid nitrogen (1100), chemicals and etc. (1600), data recording media (1100), minicomputer supplies (750)

\$ 7,150

TRAVEL

DOMESTIC 2 East Coast and 2 West Coast trips

\$ 1,400

FOREIGN

\$ 0

PATIENT COSTS (See instructions)

\$ 0

ALTERATIONS AND RENOVATIONS

\$ 0

OTHER EXPENSES (Itemize)

Telephone (office & data) - 1800
Terminal & communication equipment lease - 5240
Publications, etc. - 1800

\$ 8,840

TOTAL DIRECT COST (Enter on Page 1, Item 10)

\$ 240,962

INDIRECT
COST

(See Instructions)

47 % S&W*
47 % NTDC

*If this is a special rate (e.g. off-site), explain.

Date of DHEW Agreement:

July 30, 1973

☐ Not Requested☐ Under negotiation with:

SECTION III

**SECTION III—FISCAL DATA FOR
CURRENT BUDGET PERIOD**

(USUALLY 12 MONTHS)

FROM

5/1/74

THROUGH

7/31/75

GRANT NUMBER

R 24 RR00612-05A1

The following pertains to your CURRENT PHS budget. Do not include cost sharing funds. This information in conjunction with that provided on Page 2 will be used in determining the amount of support for the NEXT budget period.

A. BUDGET CATEGORIES		CURRENT BUDGET (As approved by awarding unit) (1)	ACTUAL EXPENDITURES THRU 3/31/75 (Insert Date) (2)	ESTIMATED ADDITIONAL EXPENDITURES AND OBLIGATIONS FOR REMAINDER OF CURRENT BUDGET PERIOD (3)	TOTAL ESTIMATED EXPENDITURES AND OBLIGATIONS (Col. 2 plus Col. 3) (4)	ESTIMATED UNOBLIGATED BALANCE (Subtract Col. 4 from Col. 1) (5)
Personnel (Salaries)		194,183	122,728	71,455	194,183	0
Fringe Benefits - included in personnel (salaries)						-
Consultant Costs		-	-	-	-	-
Equipment		105,050	92,510	10,715	103,225	1,825
Supplies		12,000	4,240	2,000	6,240	5,760
TRAVEL	Domestic	2,700	242	2,500	2,742	0
	Foreign	-	-	-	-	-
Patient Costs		-	-	-	-	-
Alterations and Renovations		-	-	-	-	-
Other		10,000	15,639	1,904	17,543	(7,543)
Total Direct Costs		323,933	235,359	88,574	323,933	0
Indirect Costs (If included in award)		114,531	76,167	38,364	114,531	0
TOTALS →		\$438,464	\$311,526	\$26,938	\$428,464	\$ 0

Use space below to:

B. List all items of equipment purchased or expected to be purchased during this budget period which have a unit cost of \$1000 or more.

C. Explain any significant balance or deficit shown in any category of Column 5.

D. List all other research support for Principal Investigator by source, project title, and annual amount.

DETAILED SALARY DATA
NIH GRANT RR 612-05A1

8/1/75-7/31/76

	<u>% Effort</u>	<u>Salary</u>	<u>Fringe Benefits</u>	<u>Total</u>
PRINCIPAL INVESTIGATORS:				
C. Djerassi	10	0	0	0
J. Lederberg	10	0	0	0
E. Feigenbaum	10	2,737	520	3,257
RESEARCH ASSOCIATES				
D. Smith	100	20,179	3,834	24,013
R. Carhart	100	18,783	3,569	22,352
H. Brown	100	20,179	3,834	24,013
G. Dromey	100	18,139	3,447	21,586
A. Duffield	15	3,784	718	4,502
PROGRAMMERS:				
W. White	50	9,177	1,744	10,921
G. Jirak	100	12,418	2,360	14,778
K. Stone	50	6,440	1,224	7,664
ELECTRONICS ENGINEER:				
N. Veizades	50	11,270	2,141	13,411
ELECTRONICS TECHNICIAN:				
D. Pearson	50	7,226	1,373	8,599
SENIOR RESEARCH ASSISTANT:				
A. Wegmann	100	17,350	3,297	20,647
RESEARCH ASSISTANTS:				
M. Stefik	100	5,528	1,050	6,578
P. Friedland	100	5,528	1,050	6,578
K. Morrill	100	3,420	650	4,070
SECRETARIAL SUPPORT:				
G. Perry	50	5,693	1,083	6,775
D. Larson	50	5,581	1,061	6,642
M. Allen	10	1,206	229	1,435
TOTAL:		\$174,638	\$33,184	\$207,822

III.C.

SUMMARY OF RESOURCE FUNDING

The interdisciplinary resource-related research project is almost wholly funded by the Biotechnology Resources Branch of the NIH. Computing support is provided by the NIH-funded SUMEX computer facility at Stanford (NIH Grant #RR00785-02, Professor Joshua Lederberg, Principal Investigator).

Additional support for chemistry research related to this grant is provided by NIH Grants GM-06840 and AM-04257 (Professor Carl Djerassi, Principal Investigator).

BUDGET JUSTIFICATION

Personnel remains the same as justified in the renewal application report with the exception of the substitution of Dr. Raymond Carhart for Dr. Natesa Sridharan. Salaries are increased by 9% per year and staff benefits are computed at 18% for the period 9/74-8/75, and are increased 1% per year thereafter, based on current University projections. Other budget categories are increased by 10% per year to account for inflation.

Equipment maintenance is budgeted for the proposed stand-alone PDP-11 system under DEC contract based on current prices. Also included is a budget for maintenance of the MAT-711 system. This estimate is based on our experience with parts replacements to date. We will provide the necessary manpower because Varian cannot provide adequate service.

Supplies are budgeted in various categories based on our operating experience to date. Electronics supplies include parts necessary for maintaining our electronics and test equipment. GC supplies include carrier gases, columns, phases, syringes, septa, etc., for GC/MS operation. The liquid nitrogen is required for cold trap operation on the MAT-711. Chemicals, glassware, etc., include the various organic chemicals, glassware, apparatus, glass tubing, etc. needed to support the recording paper for the calcomp paper and pens for ion current and spectrum plotting. Mini-computer supplies include paper, magnetic tape, ribbons, spare disk cartridges, etc., for data system operation.

The travel budget covers estimated needs (2 east coast and 2 west coast) trips for attending related professional meetings and interfacing potential program users nationally. No foreign travel is budgeted.

The "Other" budget includes operating telephone, office supplies, postage, reproduction, etc., support necessary for this project based on our previous experience. Terminal rental covers four terminals to be distributed among the MS laboratory, the Computer Science Department, and J. Lederberg's laboratory.

IV. DETAILED DESCRIPTION OF RESOURCE PROJECTS

Projects using the structure elucidation tools developed under this resource related research grant are listed in the Description of Progress (section II-A).